The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception

Raphaël Laurent, Marie-Lou Barnaud, Jean-Luc Schwartz, Pierre Bessière, Julien Diard

The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception

Raphaël Laurent, Marie-Lou Barnaud, Jean-Luc Schwartz

GIPSA-Lab

Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France

CNRS, GIPSA-Lab, F-38000 Grenoble, France

Pierre Bessière

CNRS, Sorbonne Universités - Université Pierre et Marie Curie

Institut des Systèmes Intelligents et de Robotique, Paris, France

Julien Diard

Laboratoire de Psychologie et NeuroCognition

Univ. Grenoble Alpes, LPNC, F-38000 Grenoble, France

CNRS, LPNC, F-38000 Grenoble, France

Authors' Note

The research leading to these results received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s", J.-L. Schwartz PI). This work includes research conducted as part of Raphaël Laurent's PhD thesis, defended at the Université de Grenoble on October, 8th, 2014 (manuscript in French: "COSMO : un modèle bayésien des interactions sensori-motrices dans la perception de la parole"). Portions of this work were also presented at the 9th International Seminar on Speech Production (ISSP 11) in Montréal, Canada (June 20-23, 2011), at the 14th Annual Conference of the International

Speech Communication Association (Interspeech 2013) in Lyon, France (August 25-29, 2013) and at the Workshop on Infant Language Development (WILD 15) in Stockholm, Sweden (June 10-12, 2015). The authors wish to thank Clément Moulin-Frier and Louis-Jean Boë for support and inspiration. Correspondence concerning this paper should be addressed to J.-L. Schwartz, Université Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France; CNRS, GIPSA-Lab, F-38000 Grenoble, France; 11 Rue des Mathématiques, 38400 Saint-Martin-d'Hères, France. Email: Jean-Luc.Schwartz@gipsa-lab.grenoble-inp.fr, Phone: (+33)4 76 57 47 12, Fax: (+33)4 76 57 47 10.

Abstract

There is a consensus concerning the view that both auditory and motor representations intervene in the perceptual processing of speech units. However, the question of the functional role of each of these systems remains seldom addressed and poorly understood. We capitalized on the formal framework of Bayesian Programming to develop COSMO (Communicating Objects using Sensory-Motor Operations), an integrative model that allows principled comparisons of purely motor or purely auditory implementations of a speech perception task and tests the gain of efficiency provided by their Bayesian fusion.

Here, we show three main results. (i) In a set of precisely defined "perfect conditions", auditory and motor theories of speech perception are indistinguishable. (ii) When a learning process that mimics speech development is introduced into COSMO, it departs from these perfect conditions. Then auditory recognition becomes more efficient than motor recognition in dealing with learned stimuli, while motor recognition is more efficient in adverse conditions. We interpret this result as a general "auditory-narrowband vs. motor-wideband" property. (iii) Simulations of plosive-vowel syllable recognition reveal possible cues from motor recognition for the invariant specification of the place of plosive articulation in context, that are lacking in the auditory pathway. This provides COSMO with a second property, where auditory cues would be more efficient for vowel decoding and motor cues for plosive articulation decoding. These simulations provide several predictions, which are in good agreement with experimental data and suggest that there is natural complementarity between auditory and motor processing within a perceptuo-motor theory of speech perception.

Keywords: Speech perception, computational modeling, sensory-motor interactions, adverse conditions, plosive invariance

The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception

On the functional role of auditory vs. motor systems in speech perception

Since the introduction in the 1960s of the so-called Motor Theory of Speech Perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), it is striking to remark how the debate pertaining to auditory and motor theories of speech communication has evolved. There were basically two main periods of reasoning.

The arguments from the 1960s to the 1980s mainly derived from experimental phonetics and what would now be called laboratory phonology. These were basically focused on functional questions. Auditory and motor theories were discussed according to their respective abilities to deal with the question of invariance (see an extensive review by Perkell & Klatt, 1986). Invariants were thought to exist somewhere in the acoustic signal, providing a key for abstract and categorical phonologic units from the continuous and physical substance of phonetics. The debate concerned the nature of these invariants, be this auditory or motor (see reviews of functional arguments in favor of auditory theories e.g. Diehl, Lotto, & Holt, 2004; Kingston & Diehl, 1994; Kluender 1994; Lotto 2000; Massaro & Oden 1980; Nearey, 1990; or in favor of motor invariance in Liberman et al., 1967; Liberman & Mattingly, 1985, 1989; Liberman & Whalen, 2000; and a review in Galantucci, Fowler, & Turvey, 2006).

Since the 1990s, the arguments have evolved progressively towards experimental data provided by cognitive neuroscience. With the discovery of mirror neurons (Rizzolatti, Fadiga, Gallese, & Fogassi, 1996a) and the proposal of a “mirror system” in the human perception of complex actions (Grafton, Arbib, Fadiga, & Rizzolatti, 1996; Iacoboni et al., 1999; Rizzolatti et al., 1996b), neurophysiological and behavioral experimental data made it progressively clear that the motor system plays a role in speech perception (see a recent detailed review in

Skipper, Devlin & Lametti, 2017). Evidence emerged in two steps. Firstly, neuroanatomical studies repeatedly showed that parietal and frontal brain areas associated with speech production were consistently stimulated upon speech perception tasks (e.g. Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Pulvermüller et al., 2006; Watkins, Strafella, & Paus, 2003; Wilson, Saygin, Sereno, & Iacoboni, 2004). This was particularly shown in non-standard conditions involving noise (Binder, Liebenthal, Possing, Medler, &Ward, 2004; Zekveld, Heslenfeld, Festen, & Schoonhoven, 2006), non-native stimuli (Callan, Callan, & Jones, 2014; Callan, Jones, Callan, & Akahane-Yamada, 2004; Wilson & Iacoboni, 2006), or conflicting audiovisual inputs (Jones & Callan, 2003; Ojanen et al., 2005; Skipper, van Wassenhove, Nusbaum, & Small, 2007). Secondly, behavioral studies looked for a causal role of motor areas in speech perception by altering or modulating the potential efficiency of speech motor centers, by Transcranial Magnetic Stimulation (TMS), repeated TMS or motor perturbations. Such studies have shown small but consistent perceptual effects in categorization or discrimination of speech stimuli, in ambiguous or noisy conditions (e.g., d'Ausilio et al., 2009; d'Ausilio, Bufalari, Salmas, & Fadiga ,2012; Grabski, Tremblay, Gracco, Girin, & Sato, 2013; Ito, Tiede, & Ostry, 2009; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Möttönen, Dutton, & Watkins, 2013; Möttönen & Watkins, 2009; Rogers, Möttönen, Boyles, & Watkins, 2014; Sato, Tremblay, & Gracco, 2009; Sato et al., 2011; Shiller, Sato, Gracco, & Baum, 2009).

In this context, the strong "auditory" vs. "motor" controversy about invariance at the crossroads of phonetics and phonology that prevailed until the end of the 1980s was almost completely replaced since the beginning of the 1990s by an integrative view from cognitive neuroscience, assuming that the motor and auditory systems collaborate in speech perception. This has the merits of taking into account new experimental insights, but its drawback is that the question of the respective functions of sensory and motor systems has almost completely

disappeared from the literature. However, if both auditory and motor processes do intervene in speech perception [(1)], the potential specificity and complementarity of these two systems within a perceptuo-motor speech perception architecture becomes essential. How could it be useful for speech perception to capitalize on two different systems? How could the motor system be more helpful in adverse conditions? What specific aspects of computation, for what kind of information extraction, are respectively implemented by the motor and auditory (if not visual or somatosensory) components of the speech perception system?

These are the questions we address in the theoretical framework of the "Perception-for-Action-Control Theory" (PACT). PACT is a perceptuo-motor theory of speech perception, connecting perceptual shaping and motor procedural knowledge in a principled way, in speech multisensory processing within the human brain (Schwartz, Basirat, Ménard, & Sato, 2012a; Schwartz, Boë, & Abry, 2007). PACT considers that perceptual knowledge is involved in both speech comprehension and speech control, in a communicative process. The communication unit through which parity may be achieved, is neither a sound, nor a gesture, but a perceptually-shaped gesture, that is a perceptuo-motor unit characterized both by its articulatory coherence, provided by its gestural nature and its perceptual value, necessary for function. Motor processes could be associated with multisensory processes through audio-visuo-motor binding, enabling a better extraction of adequate cues for further categorization processes (Basirat, Schwartz, & Sato, 2012; see also Skipper, van Wassenhove, Nusbaum & Small, 2007). Furthermore, perceptual categorization would benefit from motor information in addition to auditory and possibly visual clues. This would, improve variability processing and the extraction of invariance (Schwartz, Abry, Boë, & Cathiard, 2002; Schwartz et al., 2007, 2012a).

In PACT, it is also acknowledged that perception and action are co-structured in the course of speech development, which involves both producing and perceiving speech items.

The schedule of perceptuo-motor development in the first few years of age is important in this context, and seems to incorporate several major steps (Kuhl, 2004; Kuhl et al., 2008). First, auditory processes mature, enabling categorization of many phonetic contrasts almost from birth (e.g., Bertoncini, Bijeljac-Babic, Blumstein, & Mehler, 1987; Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Jusczyk & Derrah, 1987), with an early focus on the sounds of the infant's language. This can be as early as 6 months old for vowels (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992) and 10 months old for consonants (Werker & Tees, 1984). Motor processes evolve later and more slowly, beginning by articulatory exploration of the possible vocal repertoire, with canonical babbling at around 7 months of age (Davis, MacNeilage, & Matyear, 2002; MacNeilage, 1998). This continues with a later focus on the sounds of the phonological system from the end of the first year and through the following ones. Importantly, canonical babbling, sometimes considered as a purely endogenous process enabling infants to extensively explore the possibilities of their vocal tracts, seems to be influenced since its very beginning by the language heard in the surrounding environment. Such "babbling drift" has been displayed in a number of experiments concerning vowel formants, consonant-vowel associations and prosodic schemes (e. g. de Boysson-Bardies, 1993; de Boysson-Bardies, Hallé, Sagart, & Durant, 1989; de Boysson-Bardies, Sagart, & Durant, 1984).

Auditory perception is hence mature and focused before orofacial control occurs. Furthermore, the connection between the speech perception system and the motor system through the parieto-frontal dorsal pathway in the cortex does not seem to be completely mature at birth, but rather evolves throughout the first year of life (Dehaene-Lambertz, Dehaene, & Hertz-Pannier, 2002; Dehaene-Lambertz et al., 2006; vs. Kuhl, Ramírez, Bosseler, Lotus Lin, & Imada, 2014; Imada et al., 2006). Consequently, motor information would not be mature, focused or fully available for perception until the end of the first year.

157

158 **Computational models of auditory vs. motor theories of speech perception**

159 In the context of debates concerning the potential role of auditory and motor systems
160 in speech perception, computational models are likely to shed light on them by enabling
161 quantitative evaluation of some of the theoretical arguments in relation to experimental data.
162 There are already many computational models of auditory theories of speech perception.
163 Many, if not all of acoustic speech recognition systems can be construed as such, as they
164 involve the best statistical analyses of the acoustic content of large speech corpora for speech
165 understanding (see recent reviews in e.g. Hinton et al., 2012; Huang & Deng, 2010). They
166 also often incorporate more or less sophisticated computational models of the auditory
167 analysis of acoustic stimuli in the human brain (e.g. Hermansky, 1998; Deng, 1999). Auditory
168 theory models also include computational psycholinguistic models of cognitive speech
169 processing (e.g. *Trace*: McClelland & Elman, 1986; the *Distributed Cohort Model*: Gaskell &
170 Marslen-Wilson, 1997; *Parsyn*: Luce, Goldinger, Auer & Vitevitch, 2000; see also
171 Scharenborg, Norris, Ten Bosch & McQueen, 2005).

172 A widespread mathematical framework, in this domain, is probabilistic modeling,
173 where a generative, predictive model associates probable acoustic signals with linguistic
174 categories. Then, perception is cast as a categorization process, in which Bayes theorem is
175 used to infer the most likely linguistic category given some acoustic stimulus:

176 $$P([O = o_i]|S) = \frac{P(S|[O=o_i])P([O=o_i])}{\sum_j P(S|[O=o_j])P([O=o_j])},$$

177 where $P(S|[O = o_i])$ expresses the probability distribution of acoustic cues for a given
178 category and $P([O = o_i])$ defines prior probabilities of each category.

179 The origin of such models can be traced back, historically, to Signal Detection Theory
180 (Tanner & Swets, 1954; Green & Swets, 1966; more recent references include Dayan &
181 Abbott, 2001; Rouder & Lu, 2005) and its multi-dimensional generalization, the General

Recognition Theory (Ashby & Townsend, 1986; Ashby & Perrin, 1988). Recent years saw the resurgence and spread of Bayesian models of speech perception, that consider the categorization process above to model the optimal, ideal acoustic (or possibly visual, see footnote 1) information processing system, without any reference to motor processes. Such "optimal" models include ideal listener models (Feldman, Griffiths & Morgan, 2009; Sonderegger & Yu, 2010) and ideal adapter models (Clayards, Aslin, Tanenhaus & Jacobs, 2007; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Kleinschmidt & Jaeger, 2011, 2015), where categories are either learned in a batch manner (De Boer & Kuhl, 2003; Dillon, Dunbar & Idsardi, 2013) or acquired and adapted incrementally (McMurray, Aslin & Toscano, 2009; Vallabha, McClelland, Pons, Werker & Amano, 2007). In this trend, many extensions were proposed, for instance for dealing with multiple cues (Toscano & McMurray, 2008, 2010) or with higher-level structures above phonemic representations (e.g., words, syllables), in hierarchical models (Norris & McQueen, 2008; Feldman, Griffiths & Morgan, 2009b; Kiebel, Daunizeau, & Friston, 2009; Feldman, Griffiths, Goldwater & Morgan, 2013).

In comparison, there are many less computational motor theory models, basically because of the lack of easily available articulatory or motor data required for the training of such models. A few automatic speech recognition systems attempt to introduce articulatory data into their statistical processes (e.g. Deng & Ma, 2000; Deng, Ramsay, & Sun, 1997; Frankel, Richmond, King & Taylor, 2000; Sun & Deng, 2002) and a recent series of machine learning models based on artificial neural networks and applied to articulatory data recorded through electromagnetic articulography aimed to show the efficiency of articulatory inputs for phonetic decoding (e.g. Canevari, Badino, d'Ausilio, Fadiga, & Metta, 2013; Castellini et al., 2011). A few variants of Bayesian models of speech perception, because they consider computations of the speaker's intentions through motor inversion, can be construed as involving motor knowledge during perception, although it was not their initial purpose, as

they were developed instead to model perceptual magnet effects (Feldman & Griffiths, 2007; Feldman, Griffiths & Morgan, 2009). Other authors attempted to develop formal models without real articulatory ground truth data to evaluate the possibility of implementing motor or perceptuo-motor theories of speech perception (e.g. Kröger, Kannampuzha, & Kaufmann, 2014; Kröger, Kannampuzha, & Neuschaefer-Rube, 2009; Moore, 2007).

However, while all these developments basically aim to demonstrate that articulatory or motor speech decoding is indeed feasible and potentially efficient, none of this research attempts to really evaluate why and how motor information could be relevant for speech decoding. Furthermore, it is always difficult in these models to precisely disentangle what comes from the distribution of articulatory information and what comes from specific choices in the computational implementation.

This is why the Bayesian implementation of an instance of motor theory (implementing motor decoding in a Bayesian framework) or perceptuo-motor theory (including the fusion of auditory and motor information) could enable the functional role of the motor system to be assessed more clearly and rigorously. This is the objective of the COSMO model (Communicating Objects using Sensory-Motor Operations) that will be presented in the next section.

COSMO, a Bayesian computational framework for assessing the functional role of auditory vs. motor systems

To attempt to better understand the function of motor information in speech perception, we have developed over recent years a computational Bayesian framework called COSMO. This model enables auditory, motor and perceptuo-motor theories of speech communication to be implemented and compared in a coherent set of simple probabilistic equations and processes, based on Bayesian modeling. COSMO was initially developed to

deal with the emergence of sound systems in human languages (Moulin-Frier, Diard, Schwartz, & Bessière, 2015; Moulin-Frier, Schwartz, Diard, & Bessière, 2011) and was then adapted to the study of speech perception in adverse conditions (Moulin-Frier, Laurent, Bessière, Schwartz, & Diard, 2012). The present article greatly expands the initial study of Moulin-Frier et al. (2012), which attempted to clearly assess when and how motor information could be useful for phonetic decoding.

A first part will present the COSMO model, together with an initial crucial result we obtained with COSMO, which we called the indistinguishability theorem. This theorem shows that in a set of precisely defined "perfect conditions", auditory and motor theories of speech perception are indistinguishable (Moulin-Frier et al., 2012). We will present this theorem in detail, since it is of great theoretical importance, providing a landmark for any further comparison of auditory and motor models of speech perception.

Indeed, distinguishing the functional roles of auditory and motor systems for speech perception can only be achieved by departing from these perfect conditions. This can occur in one of two major ways, providing the two major contributions of the present paper to the subject.

Firstly, the learning process can differentiate the auditory and motor systems. We claim that these two systems evolve differently during learning. The auditory system could focus rapidly and precisely on the set of learning stimuli provided by the environment, leading to a system finely tuned to this learning set. This would provide the auditory system with a "narrow-band" specificity with respect to the learning data. In contrast, the motor system would "wander" more through the sensory-motor space during its exploration stage, because of the complexity of the task at hand. Hence it would evolve more slowly and focus less efficiently on the set of learning stimuli provided by the environment, in agreement with the developmental timeline described previously. However, it would be able to process a wider

set of stimuli thanks to the "wandering" phenomenon. This would provide the motor system with a "wide-band" specificity, making it poorer for learned stimuli, but better at generalizing about adverse conditions involving unlearned stimuli. This will be developed in Part 2, together with two predictions associated with this "auditory-narrow, motor-wide" property, that will be compared to the available experimental data.

Secondly, the two systems can be differentiated in terms of the nature and complexity of their internal representations, possibly leading to different processing of variability of the phonological units. Considering simulations of the recognition of plosive-vowel sequences, we explore the assumption that motor recognition might provide clues as to the invariant specification of the place of articulation of plosives in context, which is lacking in the auditory pathway, while the auditory categorization of vowels would be more straightforward than its motor counterpart. Altogether, this suggests that there should be a natural complementarity between auditory and motor systems within a perceptuo-motor theory of speech communication. This will be developed in Part 3, together with two other predictions that will be discussed in light of available experimental data.

Following the important simulation contributions and predictions, we will end this paper with a review of some major perspectives and challenges associated with the development of COSMO, in relation to cognitive processes involved in speech communication.

Part 1 – COSMO and the indistinguishability theorem

In this first part we will introduce the two major pieces of our computational framework. Firstly, we will present and describe COSMO together with its mathematical specification and the way it enables modeling of auditory, motor or perceptuo-motor theories of speech perception. Secondly, we will derive the indistinguishability theorem, already published by

Moulin-Frier et al. (2012), but which will be explained more precisely in the present paper. This will provide us with the crucial landmark that will serve for further simulations presented in parts 2 and 3.

The COSMO model

COSMO stems from the analysis of spoken communication, which can be broken down into a minimal set of variables, with a high level of abstraction.

FIGURE 1 ABOUT HERE

As is shown in the upper part of Figure 1, to communicate about an object O_S, the *Speaker* Agent performs a motor gesture M resulting in a sensory input S from which the *Listener* Agent retrieves the object O_L [(2)]. The variable C^{Env} is a Boolean variable assessing the communication success: it is *True* when $O_S = O_L$.

Our work is based on the hypothesis that the Communicating Agent, is able to act both as a speaker and a listener, and has internal representations of the whole communication loop, as shown in the lower part of Figure 1. Consequently, the Communicating Agent model is made up of (i) a *motor system* associating motor representations M to the object O_S to be produced and of (ii) a *sensory system* associating the perceived object O_L to the sensory representation S, which are linked by (iii) a *sensory-motor system* that allows the consequences of motor commands M in terms of sensory inputs S to be predicted. At this stage, M, S and O are still generic variables, in order not to lose generality. They will be instantiated for experiments and made more precise later in this paper.

The model contains two different variables O_S and O_L, one for the intention of the speaker, the other for the perception of the listener. They are also useful to avoid a directed loop from a single variable O to itself through motor and perceptual variables M and S. Indeed, such loops are not compatible with straightforward application of Bayes theorem;

duplication of the variables is a classic solution to circumvent this technical problem (e.g. state variables are replicated over time for temporal series models). Such duplication between phonological codes for production and perception, linked by conversion mechanisms, is compatible with neuropsychological data (Jacquemot, Dupoux, & Bachoud-Lévi, 2007). Coherence between variables O_S and O_L is imposed by (iv) a Boolean variable C when it is set to *True*. The variable C can also be conceived as the internalization of the C^{Env} variable assessing communication success.

This Communicating Agent model has a name, COSMO, that also happens to recall the model variables. COSMO is formally defined within the framework of Bayesian Programming (Bessière, Laugier, & Siegwart, 2008; Bessière, Mazer, Ahuactzin-Larios, & Mekhnacha, 2013; Lebeltel, Bessière, Diard, & Mazer, 2004) by probability distributions, which encode the subjective knowledge that the agent has about the relations between its internal representations. The COSMO model is thus defined by one mathematical object, the joint probability distribution over all its variables: $P(C\ O_L\ S\ M\ O_S)$. It contains all the information the agent has about its internal variables and can be shown to be sufficient to perform any inference task about these variables, whatever the form. In other words, any conditional probability over some of these variables, knowing some others, can be computed from the joint probability distribution. We chose to decompose and simplify this joint probability distribution as follows (Moulin-Frier et al., 2012):

$$P(C\ O_L\ S\ M\ O_S) = \underbrace{P(O_S)}_{prior}\ \underbrace{P(M \mid O_S)}_{\substack{motor \\ repertoire}}\ \underbrace{P(S \mid M)}_{\substack{forward \\ model}}\ \underbrace{P(O_L \mid S)}_{\substack{sensory \\ classifier}}\ \underbrace{P(C \mid O_S\ O_L)}_{\substack{communication \\ success}} . \quad (1)$$

Various tasks can then be carried out by asking questions to the model, by computing conditional probability distributions of the form $P(SEARCHED \mid OBSERVATIONS)$: What is the probability distribution over the $SEARCHED$ variables, knowing the value of some $OBSERVATIONS$? In the COSMO framework, a speech production task amounts to

computing a conditional distribution of the form $P(M \mid O)$: What is the probability distribution over motor commands M corresponding to the object O to be communicated? A speech perception task amounts to computing a conditional distribution of the form $P(O \mid S)$: What is the probability distribution over perceived objects O, given the sensory input S?

Within the framework of *COSMO*, these questions can be instantiated in three different ways: (i) by replacing O by O_S we implement a motor theory focused on the speaker's perspective, (ii) by replacing O by O_L we implement an auditory theory focused on the listener's perspective, (iii) by indifferently using either O_S or O_L and by further conditioning the computed distribution with the constraint $C = True$, we implement a perceptuo-motor theory that ensures the coherence of both representations. This is the equivalent of explicitly setting the communication success as a goal of the task considered. Bayesian inference provides a way to compute the conditional probability distributions corresponding to all these tasks from the joint probability distribution that defines the *COSMO* model (Equation (1)). Figure 2 shows the results of these computations and how they can be interpreted. We now explain further the results of these Bayesian inferences, focusing on the speech perception task.

FIGURE 2 ABOUT HERE

As can be seen in Figure 2, the implementation of an auditory theory of perception consists of a direct computation of $P(O \mid S) = P(O_L \mid S)$, with no intervention of motor variables. This is consistent with classical proposals about auditory theories, which deny the role of motor knowledge in speech perception and consider that it is based exclusively on the set of auditory processing and categorization mechanisms available in the human brain (e.g. Diehl et al., 2004). We note that this portion of our model is equivalent to many preceding models of acoustic categorization, including the Ideal Adapter model of Kleinschmidt & Jaeger (2015) mentioned previously.

Likewise, the implementation of a motor theory of perception, i.e., computing $P(O \mid S) = P(O_S \mid S) \propto \sum_M (P(M \mid O_S)\, P(S \mid M))$ (Feldman & Griffiths, 2007; Feldman, Griffiths & Morgan, 2009; Moulin-Frier et al., 2012), is consistent with the view that speech perception occurs by retrieving the intended motor gestures of the speaker (Liberman & Mattingly, 1985). Indeed, the motor variable M now plays a role in the inference. The deterministic two-stage process posited by motor theories begins with the retrieval of M from S through an inverse model, which is followed by the categorization process estimating O_S from M through a motor decoder. In the Bayesian framework, these are replaced by the computation of the sum over the possible values of the variable M, weighted by the probability that they have the sensory consequence S and by the probability that they are associated with O_S the considered object. This is a Bayesian analogue to analysis-by-synthesis (Halle & Stevens, 1959; Stevens & Halle, 1967; see a review in Bever & Poeppel, 2010). The deterministic two-stage process, firstly with motor-to-sensory inversion and secondly with motor decoding, is an approximation of the summation over all M values.

Finally, the implementation of a perceptuo-motor theory of perception consists simply of a mere Bayesian fusion of the predictions of the sensory and motor categorization processes: $P([O = o] \mid S\, [C = True]) \propto P([O_L = o] \mid S)\ \sum_M (P(M \mid [O_S = o])\, P(S \mid M))$.

Indistinguishability of auditory and motor theories in perfect conditions of learning and communication

Although purely sensory and purely motor perceptions are described by different equations (see Figure 2), it can be proven that if three hypotheses defining a set of "perfect conditions" of learning are verified, the motor and auditory theories of perception make exactly the same predictions. Therefore, these cannot be distinguished empirically. This demonstration has been presented previously (Moulin-Frier et al., 2012), but in a less explicit

formulation. We will present it here again in detail, in its more rigorous form, with three hypotheses instead of the two used previously.

FIGURE 3 ABOUT HERE

We consider a supervised learning scenario, shown in Figure 3, which features Learning Agents and a Master Agent, each described as a COSMO agent. To distinguish their variables, superscripts are added and variables become O_S^{Ag}, O_S^{Master}, M^{Ag}, M^{Master}, etc.

In the learning scenario, the Learning Agent is provided by the Master Agent with the following <object, stimulus> pairs. The Master Agent uniformly selects O_S^{Master} objects, draws corresponding M^{Master} motor commands according to the production model $P(M^{Master} \mid O_S^{Master})$, which are then transformed by the environment modeled by $P(S^{Ag} \mid M^{Master})$ and result in sensory S^{Ag} inputs. Furthermore, the variable C^{Env}, which ensures coherence between the O_S^{Master} and O_L^{Ag} objects, implements a shared attention mechanism, e.g. deixis, which allows the Learning Agent to retrieve the right objects (O_L^{Ag}) from the Master to associate with the S^{Ag} stimuli in its sensory classifiers $P(O_L^{Ag} \mid S^{Ag})$. The Learning Agent builds its sensory classifier through successive random draws, which are mathematically expressed by the following approximation:

$$P(O_L^{Ag} \mid S^{Ag}) \approx \sum_M (P(M^{Master} \mid O_S^{Master}).P(S^{Ag} \mid M^{Master})). \quad (2)$$

In this equation, the sign ≈ expresses the fact that the set of learning stimuli (right part of the equation) has to be learned in some way from the $P(O_L^{Ag} \mid S^{Ag})$ distribution (left part of the equation).

We now define the three hypotheses used in this approach and prove that their conjunction ensures the indistinguishability of the motor and auditory theories of speech perception.

i. H1 (the "perfect sensory learning hypothesis"): the sensory classifier is

perfectly learned from the Master's productions, i.e. $P(O_L^{Ag} \mid S^{Ag}) = \sum_M (P(M^{Master} \mid O_S^{Master}).P(S^{Ag} \mid M^{Master}))$. By replacing the operator ≈ of Equation (2) by an equality operator =, H1 explicitly states that the sensory classifier $P(O_L^{Ag} \mid S^{Ag})$ learned by the agent perfectly encodes all the information expressed by the combination of the probability distributions $P(M^{Master} \mid O_S^{Master})$ and $P(S^{Ag} \mid M^{Master})$. These describe the way the Master performs its motor gestures and the way they are transformed by the environment.

ii. H2 (the "perfect motor learning hypothesis"): the motor repertoire of the agent is identical to that of the Master, i.e. $P(M^{Ag} \mid O_S^{Ag}) = P(M^{Master} \mid O_S^{Master})$.

iii. H3 (the "perfect sensory-motor learning hypothesis"): the agent's sensory-motor system perfectly encodes the properties of the transformation performed by the environment during the learning process, i.e. $P(S^{Ag} \mid M^{Ag}) = P(S^{Ag} \mid M^{Master})$.

The indistinguishability theorem states that if H1, H2 and H3 hold, then the motor and sensory instantiations of the speech perception task are indistinguishable.

The proof is straightforward. Starting from Equation (2), which states how the sensory decoder is learned along the paradigm in Figure 3, hypothesis H1 enables the learning operator ≈ to be replaced by an equality operator =, while hypotheses H2 and H3 enable the two terms on the right hand side of Equation (2) to be replaced by $P(M^{Ag} \mid O_S^{Ag})$ and $P(S^{Ag} \mid M^{Ag})$, respectively, which yields:

$$P(O_L^{Ag} \mid S^{Ag}) = \sum_M \big(P(M^{Ag} \mid O_S^{Ag}).P(S^{Ag} \mid M^{Ag})\big). \quad (3)$$

The right hand side of Equation (3) has now become the expression of the motor instantiation of the speech perception task, while the left hand side is the expression of the perception task instantiated within the framework of the auditory theory (see Figure 2).

Therefore, if these three hypotheses are verified within a set of "perfect conditions" for learning, the sensory and motor models rely on the same information and make the same predictions. They are thus indistinguishable, whatever the test conditions might be.

When the indistinguishability theorem is satisfied, information encoded in the motor and sensory pathways is redundant. This shows that even when two theories or models are seemingly different – as the auditory and motor theories of speech perception appear to be – they may be identical with respect to the computation they perform (as conceptualized by Marr, 1982, in his three-level framework, in which the same computational task can be carried out by algorithmic models with different representations; see also Laurent, Schwartz, Bessière, & Diard, 2013).

Similar arguments are sometimes invoked in papers about auditory theories (e.g. Diehl et al, 2004, p. 168: "listeners do not recover gestures, but they do perceive the acoustic consequences of gestures. Any regularities of speech production (e.g., context dependencies) will be reflected in the acoustic signal and, through general mechanisms of perceptual learning, listeners come to make use of the acoustic correlates of these production regularities in judging the phonemic content of speech signals"). The indistinguishability theorem provides a theoretical basis based on Bayesian modeling to explain such more or less intuitive claims. More importantly, it suggests that what should drive our understanding of the respective roles of the auditory vs. motor systems in speech perception is related to what we are able to learn about them in the course of speech development.

Understanding the potential role and complementarity of the sensory and motor recognition processes requires departing from the perfect conditions defined previously. Given the structure of the motor and sensory models, the possible differences between their predictions of perception tasks are strongly dependent on the information they encode, i.e., on how they were learned.

The next parts of this article will introduce two sets of simulations providing two directions in which auditory and motor theories depart from each other. Furthermore, some fundamental sources of functional complementarity will be displayed.

Part 2 – The "auditory-narrow, motor-wide" framework for speech perception

In this part, we will focus on a *generic property* of COSMO that we consider largely independent from the specific implementation choices and that refers to structural aspects of the way auditory vs. motor decoding can be modeled in a Bayesian framework. This generic property generates a natural complementarity between auditory and motor decoding processes, that we summarize by the so-called "auditory-narrow, motor-wide" framework. Finally we discuss the relationship between simulations and experimental data for speech perception development and speech processing in noise.

The sensory branch is narrow-band, the motor branch is wide-band: simulations within a simplified one-dimensional sensory-motor space

Equations in Figure 2 defining motor vs. sensory categorization show a major structural difference between the two processes. While sensory perception implements a direct association between the sensory input S and the perceived object O_L, motor perception appears to be more complex. Indeed the pathway from S to O_S involves motor information, M. This suggests that motor recognition might require more time or cognitive resources before convergence in the learning process, compared to sensory recognition. A possible consequence is that the sensory system should be able to focus more rapidly and efficiently on the set of exogenous learning stimuli provided by the environment, while the motor system "wanders" through the sensory-motor space and endogenously explores regions, possibly

different ones from the exogenous input. This would provide the sensory and motor systems with what we have called a "narrow-band" vs. "wide-band" specificity with respect to the learning data. The latter would be less efficient for learned stimuli, but would function better in adverse conditions involving unlearned stimuli.

This is what we set out to demonstrate, on a highly simplified theoretical framework based on 1-D motor and sensory variables linked by a sigmoid transformation. In this section the variables of the *COSMO* model are constrained to be very simple and are instantiated as follows: M and S are 1-D and discrete (with values regularly distributed between -15 and 15), while O_S and O_L both denote two possible objects o_1 and o_2. The Master Agent and the Learning Agent correspond to two different instances of the *COSMO* model with the same parametric forms (mostly Gaussian probability distributions) mathematically encoding the knowledge stored in the models. The two types of agent only differ by the values of the parameters of these parametric forms (for instance, means and standard deviations of the Gaussian probability distributions). We consider a supervised learning situation, where the parameters of the Master Agent and of the motor-to-sensory transformation performed by the simulated environment are fixed and the Learning Agent determines values for its parameters of internal representations through interactions with the Master according to the supervised learning scenario shown in Figure 3. We now describe the probability distribution forms and the parameters that are constant throughout learning. The prior objects $P(O_S)$ for both types of agent are set as uniform probability distributions; objects o_1 and o_2 are produced by the Master with the same frequency and the Learning Agent has no prior knowledge of the frequency of object apparition.

For both types of agent, motor repertoire probability distributions $P(M \mid O_S)$ are encoded as Gaussian probability distributions. For instance, to select a motor command corresponding to object o_1, the Master Agent draws a value of M^{Master} according to the

probability distribution $P(M^{Master} \mid [O_S^{Master} = o_1]) = Gauss(\mu_1^M, \sigma_1^M)$, where the mean value μ_1^M of the Gaussian probability distribution corresponds to a prototypic motor gesture and the standard deviation σ_1^M quantifies the variability of the Master Agent's production. In the Master Agent model, we set $\mu_1^M = -5$, $\mu_2^M = 5$ and $\sigma_1^M = \sigma_2^M = 1$ (see Figure 4, bottom plot).

The motor-to-sensory transformation $P(S \mid M)$ occurring in the environment is modeled as Gaussian probability distributions. More precisely, when the Master Agent issues a motor command m, the Learning Agent receives a value of the sensory input S^{Ag} drawn according to the probability distribution $P(S^{Ag} \mid [M^{Master} = m]) = Gauss(\mu_m^S, \sigma_m^S)$, where the value $\mu_m^S = f(m)$ is given by a function f modeling the motor-to-sensory transformation and $\sigma_m^S = \sigma^{Env} = 1$ is a constant encoding the communication noise at learning time.

FIGURE 4 ABOUT HERE

Next, we consider nonlinear monotonous transformations, to keep some level of generality. Interestingly, nonlinear motor-to-sensory transformations have been exploited by the Quantal Theory of Speech (Stevens, 1972) as providing natural category boundaries for phonetic contrasts. In the Quantal Theory of Speech, it is proposed that such nonlinearities lead to the existence of articulatory plateaus, where variations in the articulation input lead essentially to no, or only small, acoustic variations. These are separated by discontinuity regions, where a small articulatory variation results in a strong acoustic jump. Stevens (1972, 1989) suggested that human languages exploit such discontinuities to set universal phonetic contrasts. This principle was confirmed in COSMO simulations of the emergence of phonological systems (Moulin-Frier et al., 2015). In the present study, we define the physical link f between the motor gestures M and their sensory consequences S as a sigmoid function $f(m) = b.\frac{\tan^{-1}(a.m)}{\tan^{-1}(a.b)}$, which is shown in Figure 4 (top left plot). Parameter a allows the slope of the sigmoid function f to be tuned and parameter b controls its range. We selected a b

value of 12, slightly lower than the M and S values of 15 and we set a to either 0.01 in a quasi-linear case, or 0.1 to obtain a nonlinear case compatible with the Quantal Theory. The corresponding probability distributions $P(S^{Ag} \mid [O_S^{Master} = o_1])$ and $P(S^{Ag} \mid [O_S^{Master} = o_2])$ are displayed in Figure 4 (top right plot). In the nonlinear case they are naturally more widely separated in sensory space than the equivalent distribution in motor space.

At this stage, we consider that the Master, along with the Learning Agent, have the same nonlinear transformation between their motor and sensory variables. A major departure from this assumption would concern differences in vocal tract shape mainly associated with age and sex. We consider that all agents are equipped with a normalization mechanism enabling them to transform sensory information provided by the Master into an S value appropriately situated in their internal sensory space. Such normalization processes exist and have been displayed since the first months of age (e.g. Kuhl, 1979, 1983; Polka, Masapollo and Ménard, 2014). Once the stimuli are transmitted by the Master to the Learning Agent and have been appropriately normalized, the nonlinear transformation is of no further use to the Master Agent. Hence, remaining differences between such transformations in the case of the Master and the Learning Agent play no role in further processing in COSMO. We will return in the discussion of Part 2 to consider how realistic normalization processes could modify the present simulations.

In the computer simulations presented below, all Gaussian probability distributions are truncated: we define a baseline value $\varepsilon = 10^{-5}$ and probability values below this threshold are set to ε; the probability distribution is normalized afterward [(3)]. This avoids cognitively implausible numerical precision of probability distributions, which would yield unwanted side effects. For instance, in classification tasks, when comparing the predictions of Gaussian models too far from their mean values, an infinitely precise model generalizes too well and behaves like an analytical model, with a precise classification frontier and abruptly changing

responses. In the present simulations, probability distributions, truncated in this manner, degenerate outside of their "competence domains" and classification responses behave according to chance when exotic stimuli are presented.

Learning in *COSMO*

Hypotheses H1, H2 and H3 are at the basis of the indistinguishability theorem, expressing unrealistic perfect learning conditions. We now add a plausible learning algorithm to the hypotheses and we will describe how the result departs from ideal learning conditions and ultimately enables sensory and motor recognition processes to be distinguished.

Learning follows the interaction paradigm introduced as Figure 3. To recapitulate the learning scenario, both the Master and Learning Agent interact in a simulated environment. Probability distributions defining the Master Agent and motor-to-sensory transformation of the environment are set constant during learning. The Master Agent provides the Learning Agent with <object, stimulus> pairs (later referred to as $< o, s >$). The Learning Agent ascertains from this data both its sensory and motor classification systems; more precisely, it identifies parameters for its sensory prototypes $P(S^{Ag} \mid O_L^{Ag})$, internal model $P(S^{Ag} \mid M^{Ag})$and motor repertoire $P(M^{Ag} \mid O_S^{Ag})$, which are all implemented using Gaussian probability distributions. To express the fact that the Learning Agent starts without any knowledge, initial states of all these Gaussian probability distributions are characterized by values of mean parameters μ at the center of their domains and initial standard deviation values σ that are large relative to the domain size. This approximates to uniform probability distributions.

To allow fair comparisons of the sensory and motor instantiations of the speech perception task, the components of the sensory and motor classification systems are learned independently and using the same data.

In the sensory recognition system based on a direct association between stimuli and objects, firstly we learn sensory prototypes of the form $P(S^{Ag} \mid [O_L^{Ag} = o]) = Gauss(\mu_o^S, \sigma_o^S)$ which correspond to each object. Learning consists of computing a Gaussian probability fit distribution $P(S^{Ag} \mid O_L^{Ag})$ from $< o, s >$ pairs. Each time the Learning Agent receives such a pair from its Master, the values of the mean μ_o^S and of the variance σ_o^S of the corresponding Gaussian probability distribution are updated accordingly. An extensive study of the dynamics of learning sensory prototypes and their effects on the resulting classifiers is provided by Kleinschmidt & Jaeger (2015); as such dynamics are not the focus of our contribution, we implement a straightforward learning procedure, where the order of learning data has no effect.

The motor recognition system exploits the same $< o, s >$ pairs to learn both components of its pathway, namely the internal model of the motor-to-sensory transformation $P(S^{Ag} \mid M^{Ag})$ and the motor repertoire $P(M^{Ag} \mid O_S^{Ag})$. We exploit a *Learning by Accommodation* algorithm, which allows learning of the two components at the same time. Importantly, this algorithm takes into account "babbling drift", i.e. the fact, presented in the Introduction, that the agent should not explore systematically and uniformly its sensory-motor space but rather should focus on regions of interest provided by the Master's sounds. The accommodation algorithm enables the Learning Agent to progressively focus on the Master's stimuli, making learning quicker and more efficient (Barnaud, Schwartz, Diard, & Bessière, 2016).

The algorithm involves a simple imitation paradigm without any error measurements. It works in the following way:

(i) The Learning Agent tries to mimic the sensory input s of the $< o, s >$ pair provided by the Master, by producing a motor command m given its current state of knowledge, the input stimulus s and the input object o. After probabilistic inference, this amounts to

randomly drawing a value for m according to the following probability distribution:

$$P(M^{Ag} \mid [S^{Ag} = s]\left[O_S^{Ag} = o\right]) = P(M^{Ag} \mid \left[O_S^{Ag} = o\right]).P([S^{Ag} = s] \mid M^{Ag}). \quad (4)$$

Equation (4) shows that the choice of motor commands m is driven by two factors: first, the need to match the stimulus s given by the Master, as predicted by the current state of knowledge encoded in the internal model $P(S^{Ag} \mid M^{Ag})$; second, the tendency to use the same motor commands that were previously associated with the object o communicated by the Master, as stored in the motor repertoire $P(M^{Ag} \mid O_S^{Ag})$.

(ii) Once selected, the motor command m is performed and has a sensory consequence s'. The Learning Agent then uses the observed correspondence of s' and m to improve the internal model by updating the parameters μ_m^S and σ_m^S of the probability distribution $P(S^{Ag} \mid [M^{Ag} = m])$. It also exploits the selected value m in its motor repertoire by updating the μ_o^M and σ_o^M parameters of the probability distribution $P(M^{Ag} \mid \left[O_S^{Ag} = o\right])$.

Therefore, the algorithm progressively refines the internal model of motor-to-sensory mapping, both with some endogenous random exploration due to inaccurate imitation in the first stages of the learning process and with a progressive focus on the learning stimuli that result in a better mapping around the regions of the stimuli provided by the Master. In parallel, the algorithm progressively anchors adequate motor gestures for each object, i.e. gestures producing sounds that correspond to the sensory distribution produced by the Master for the object.

Simulation results

Learning pace: fast and focused sensory learning vs. slow and diffuse motor learning.

FIGURE 5 ABOUT HERE

We use the evolution of entropy $H\big(P(X)\big) = -\sum_i P(X = x_i) \log P(X = x_i)$ as a

numeric indicator that quantifies how much information becomes stored in the probability distributions of the models. We compare learning speeds using the evolution of $H\left(P\left(S^{Ag} \mid O_L^{Ag}\right)\right)$, the entropy of the sensory model on the one hand and $H\left(P\left(S^{Ag} \mid O_S^{Ag}\right)\right)$, the entropy of the motor model on the other hand. $H\left(P\left(S^{Ag} \mid O_S^{Master}\right)\right)$, the entropy of the probability distribution over the stimuli produced by the Master Agent, which is constant during learning, is used as a reference. Each of these entropy values is actually a set of measurements, one for each possible object value. We therefore average them over objects and, since we have two objects in these 1-D experiments, consider $\frac{1}{2}\sum_{O_L^{Ag}} H\left(P\left(S^{Ag} \mid O_L^{Ag}\right)\right)$ for the sensory model, $\frac{1}{2}\sum_{O_S^{Ag}} H\left(P\left(S^{Ag} \mid O_S^{Ag}\right)\right)$ for the motor model and $\frac{1}{2}\sum_{O_S^{Master}} H\left(P\left(S^{Ag} \mid O_S^{Master}\right)\right)$ as the Master's reference.

The corresponding curves are displayed in Figure 5 for the two values of nonlinearity in the motor-to-sensory transformation. This Figure shows that the entropy of the sensory model converges quickly to a level close to the entropy of the stimuli produced by the Master, while the entropy of the motor model converges more slowly. In the linear case, the sensory model is able to converge to exactly the same entropy as that of the Master Agent, whereas it remains larger in the nonlinear case because the constraint that distributions $P\left(S^{Ag} \mid O_L^{Ag}\right)$ are Gaussian leads to some residual discrepancy between the models of the Master Agent and Learning Agent.

However, whatever the nonlinearity, the motor model entropy $H\left(P\left(S^{Ag} \mid O_S^{Ag}\right)\right)$ decreases more slowly than the sensory model entropy, indicating slower learning. This corresponds to our prediction that the inference process is more complex in the motor model. Hence the learning mechanism is slower and less efficient, since it "visits" portions of the sensory-motor space that are not available to the sensory recognition system.

FIGURE 6 ABOUT HERE

To support this point, let us recall that learning the motor $P\left(S^{Ag} \mid O_S^{Ag}\right)$ model consists of learning both $P(S^{Ag} \mid M^{Ag})$ and $P(M^{Ag} \mid O_S^{Ag})$. We show in Figure 6 an instance of an internal $P(S^{Ag} \mid M^{Ag})$ model in the nonlinear case learned by the agent after 20,000 iterations of learning by the accommodation algorithm. Data presented in Figure 6 first show that the shape of the motor-to-sensory transformation has been adequately learned. However, it appears that some regions are learned better than others: these regions, where the variance of the $P(S^{Ag} \mid [M^{Ag} = m])$ distribution is small, correspond to those of the sensory space where stimuli have been provided by the Master Agent. Other regions, far from the data of the learning set, have a higher variance. However, the Learning Agent has acquired global knowledge, which provides the motor learning process with what we could call a more "diffuse" character.

In this learning process, we have implemented a mechanism that naturally leads to the departure from both hypotheses H1 and H2 of the "perfect learning conditions". Since learning is intrinsically incomplete, the Learning Agent cannot fully internalize all the production abilities of the Master Agent. This results in complementarity between the sensory and motor models. While the sensory system can focus on the stimuli provided by the Master Agent and learn them quickly and efficiently, the motor system has to learn both a sensory-motor model and a motor repertoire. This more complex process is slower and less focused on the learning set, because it requires exploring an intermediate motor space. However, this can be useful for unlearned conditions as we will assess next.

Evaluation of perception: the sensory model is better in clear speech, whereas the motor model is more robust in noisy conditions.

In this section we compare the models' robustness to communication noise in an

679 evaluation experiment where the Learning Agent interacts with a Master Agent defined in the
680 same way as previously, except that after the learning phase, we introduce a test phase where
681 we vary the standard deviation $\sigma_m^S = \sigma^{Env}$ of the Gaussian probability distribution
682 $P(S^{Ag} \mid [M^{Master} = m])$, thus encoding various levels of environmental noise.

683 The Master Agent provides $< o, s >$ pairs of a given noise level and the agent
684 estimates the object o' from the stimulus *s* using either sensory recognition, i.e. by computing
685 the probability distribution $P\left(O_L^{Ag} \mid S^{Ag}\right)$, or motor recognition, i.e. by computing
686 $P\left(O_S^{Ag} \mid S^{Ag}\right)$, as defined in Figure 2. Sensory recognition is implemented as a Gaussian
687 classifier obtained by probabilistic inversion of the sensory prototypes $P\left(S^{Ag} \mid O_L^{Ag}\right)$:

688 $$P\left(O_L^{Ag} \mid S^{Ag}\right) = \frac{P\left(S^{Ag} \mid O_L^{Ag}\right)}{\sum_{O_L^{Ag}} P\left(S^{Ag} \mid O_L^{Ag}\right)} .$$

689 Motor recognition is implemented according to:

690 $P(O_S^{Ag} \mid S^{Ag}) \propto \sum_{M^{Ag}} (P(M^{Ag} \mid O_S^{Ag})\, P(S^{Ag} \mid M^{Ag}))$.

691 Comparing the values of the object intended by the Master Agent, o, and that
692 estimated by the Learning Agent, o', we compute confusion matrices and define the
693 recognition rate as the mean of their diagonal coefficients.

694 ***FIGURE 7 ABOUT HERE***

695 In Figure 7, we present the mean values of recognition rates for the linear and
696 nonlinear cases. The scores are provided at three learning stages (after learning 500, 2,000 or
697 20,000 $< o, s >$ pairs), for a range of noise degradation, from no added noise to stimuli
698 corrupted by high levels of noise (noise is indexed by variation of the σ^{Env} value).

699 First, we observe a large effect of nonlinearity on the sensory classifier, with a sharp
700 decline of performance with noise in the nonlinear case. This derives from more pointed and
701 separated probability distributions (Figure 4, top right panel). The observations that follow are
702 independent of nonlinearity.

Second, since the sensory system learns rapidly, it has already converged before 500 learning iterations and does not evolve afterwards. It provides good recognition scores without noise, with a quick degradation of performance when noise is added.

Third, in contrast the motor system appears to learn slowly. At the beginning of the learning process (top row in Figure 7), it performs very poorly and the decrease of the recognition rate as noise increases is slower than for the sensory model. When learning proceeds with more iterations, the motor system performs increasingly well, the general trend being that it becomes better than the sensory model in noisy conditions, though still remaining poorer in the absence of noise. At the last stage of the learning process (20,000 iterations) the two models give rather similar performances (we tend towards the "perfect learning conditions" of the indistinguishability theorem).

Fourth, and finally, the perceptuo-motor model implementing a Bayesian fusion of the sensory and motor recognition models according to the Equation shown in Figure 2 performs better than the two isolated models under all conditions.

FIGURE 8 ABOUT HERE

We now explore how the sensory system is more efficient in the absence of noise and how the motor system is more efficient in its presence. This is illustrated in Figure 8, where we display probability distributions for the two objects, for both motor and sensory systems. Furthermore, we show the example s_{clean} stimuli for a stimulus under normal conditions, i.e. without added noise, and s_{noise} for a stimulus in adverse conditions, i.e. with added noise.

When the "typical" s_{clean} stimulus is considered, it is close to prototypes of the motor and sensory models, i.e. to the modes of corresponding probability distributions. However, the sensory model, being of lower variance than the motor model, yields a less uncertain probability distribution categorization than the motor process. The two models correctly recognize object o_2 as the cause of the s_{clean} stimulus, but the sensory model is slightly more

certain of perception than the motor model is.

When the "noisy" stimulus (s_{noise}) is considered, it is far from prototypes of both motor and sensory models. However, the motor model, being of greater variance than the sensory model, generalizes better. Whereas for the sensory model, probability distributions quickly fall below the ε threshold we defined, yielding random categorization, the motor model is more robust and conserves categorization capabilities.

Concluding Part 2: summary and predictions

The present simulations let a major difference between auditory and motor learning appear. Auditory learning is rapid and, by definition, perfectly focused on the acoustic stimuli provided by the Master. In fact, the auditory system in COSMO is an "ideal processor" of acoustic input, as in many previous Bayesian models (e.g. Norris & McQueen, 2008; Kleinschmidt & Jaeger, 2015). However, the intrinsic limitation is provided by departures from exactly what has been learned. This is where the motor system may become relevant. Indeed, the motor system is intrinsically slower since it has a more complex inference process to deal with. It is also less well tuned to the learning corpus, because of the existence of an intermediate motor representation in the inference process. But it is this more complex learning process that supplies the possibility of wandering around stimuli and configurations that are not contained in the learning set provided by the environment. This is what makes it "wider" and hence better able to process unknown stimuli.

It is important to stress at this stage that the auditory-narrow, motor-wide hypothesis appears to be generic, i.e. intimately related to the basic COSMO structure, because of the more complex structure of the motor inference process compared with the auditory one (see Figure 2). We had to propose a number of technical and non-generic choices to perform simulations in this part of the article. These include: (i) the motor-to-sensory transformation

was presumed to be nonlinear but monotonous, (ii) the Master was considered to be physically similar to the Learning Agents, with the same nonlinear motor-to-sensory transformation, supposing that the normalization process was solved in some way, (iii) only one Master was introduced into the learning process, while an infant typically has to deal with a number of Masters to learn from in the environment.

More realistic simulations, involving: non-monotonous motor-to-sensory transformations, variations of transformations from one agent to the other, possibly in relation with normalization processes between agents with different sizes and shapes of their vocal tract, multiple Master Agents in the learning process, would basically result in a large increase in complexity of the sensory-to-motor inference process and hence in an increase toward the trend for slow and diffuse motor learning. In some sense, the 1-D simulations presented in Part 2 minimize the trend towards the auditory-narrow vs. motor-wide contrast, which is likely to be larger in a more realistic simulation with COSMO – as will be displayed in Part 3. Therefore, it can safely be claimed that the auditory-narrow motor-wide hypothesis is an intrinsic property of the COSMO structure, and probably an intrinsic characteristic of motor vs. auditory decoding in a perceptuo-motor theory of speech perception.

This property of the model generates two predictions, in the sense that two consequences follow directly from the property. These consequences were not considered during modeling; they are logically entailed by the model. These predictions are in line with already available data and observations pertaining to speech development and processing.

Prediction 1 - auditory learning should be more rapid than motor learning

A strong prediction in COSMO is that auditory learning, which typically consists of learning the sensory distributions $P(S^{Ag}/O_L{}^{Ag})$, is a simpler process than motor learning i.e. learning the motor distributions $P(M^{Ag}/O_S{}^{Ag})$. It is well-known that the auditory system is developmentally mature before the motor one, as is reviewed in the Introduction, but this is

generally only related to biological constraints. Firstly, audition begins to mature before birth, as is displayed by the sensitivity of newborns to language (Mehler, Jusczyk, Lambertz, Halsted, Bertoncini, & Amiel-Tison, 1988) or to the voice of their mother (DeCasper & Fifer, 1980). Secondly, critical periods seem to shape the course of development of speech perception and production towards the mature stage (see a recent review in Werker & Hensch, 2015). Importantly, the present simulations suggest that an additional factor could be provided by the complexity of the learning process. In this respect, it is of interest to mention that even for vowels, language tuning in production has never been described before 10 months of age (e.g. de Boysson-Bardies et al., 1989) while it occurs in speech perception as soon as 6 months of age (Kuhl et al., 1992), though infants are capable of producing vowel-like vocalizations almost since birth and display vocal imitations as early as 4 months of age (Kuhl & Meltzoff, 1996).

It would be of great interest in this discussion to attempt to correlate observed delays in the developmental schedule with some measurement of differences in the learning period or entropy reduction in a Figure such as Figure 5. However, this seems far from any reasonable prediction at the present state of possible simulations.

Prediction 2 - motor processing should be more important in adverse conditions

This is the major prediction of the auditory-narrow motor-wide hypothesis. Indeed, it is proposed as an intrinsic COSMO property that the motor system should be less efficient than the auditory system in learned conditions, while the motor system gains efficiency in unlearned ones, e.g. in noisy or adverse conditions. A likely consequence of this prediction is that the motor system should be more involved in such adverse conditions. As reported previously, this is exactly what is regularly observed for neurocognitive data, with an increased BOLD (Blood-Oxygen Level Dependent) activity in fMRI (functional Magnetic Resonance Imaging) data in motor regions for noisy (Binder et al., 2004; Zekveld et al.,

803 2006), or non-native stimuli (Callan et al., 2004, 2014; Wilson & Iacoboni, 2006). This is also
804 in line with evidence for motor perturbations seen in auditory perception only in noisy
805 conditions (e.g. d'Ausilio et al., 2012 vs. 2009), or for ambiguous stimuli around a phonetic
806 boundary (e.g. Möttönen & Watkins, 2009; Rogers et al., 2014).
807
808

Part 3 – Extracting perceptuo-motor invariance in syllabic units

While in the previous part the focus was on generic properties of the COSMO model, we now move towards non-generic properties associated with the specific way auditory and motor information is most probably distributed in a specific case of phonetic sequences made of CV syllables with a C stop consonant and a V oral vowel. This part of the article will deal with a problem that has long been considered crucial in the debate about auditory vs. motor theories, namely the invariance problem. The question of invariance has often been raised by motor theorists around the alleged lack of acoustic invariance for the plosive place of articulation, considering that in this specific case motor invariance was straightforward (see below). However, the case of vowels seems different and it has already been suggested that invariance could be of a different nature for vowel vs. plosive place of articulation, which would be auditory in one case and articulatory in the other (see Bailly, 1997; Kröger et al., 2009, 2014). Therefore, COSMO appears as a perfect framework for dealing with this question in a perceptuo-motor framework.

To address the question of auditory vs. motor invariance for vowel vs. plosive place of articulation, we will need to introduce specific knowledge and hypotheses concerning CV syllable production, perception and development. Furthermore, simulations will be carried out on a specific model of the vocal tract, VLAM (variable linear articulatory model), enabling generation of articulatory and acoustic configurations associated with CV sequences. Finally, simulations will include specific simplifications about sensory and motor variables, as well as about the learning process.

In light of this specific implementation of COSMO for syllables, which we will call COSMO-S, the question we address is: in the distribution of information for plosives and vowels, is there any potential evidence for differentiation of the motor and auditory systems in extracting phonetic invariance from acoustic stimuli? We will firstly present a literature

review on the perception and production of CV syllables in relation to the place of articulation cues. Then we will describe the vocal tract model VLAM, and the COSMO-S version of COSMO for syllable perception and production, together with the way learning is implemented in COSMO-S. Finally, we will describe simulations with COSMO-S and explore what light they might shed on the question of vowel vs. plosive place of articulation invariance.

Auditory or motor cues for vowel vs. plosive place of articulation

The question of the plosive place of articulation invariance has long been considered as a crucial test for auditory vs. motor theories of segmental invariance. On the one hand, partisans of motor theories have regularly mentioned it as a typical case, where auditory invariance was out of reach while motor invariance would be directly available (Liberman et al., 1967; Liberman & Mattingly, 1985). On the other hand, the classical objection to the motor theory is the probable complexity of the cognitive or computational implementation of the inversion process that would enable the listener to recover the proposed motor invariant from the acoustic speech input: the labial gesture for bilabials, the tongue apex gesture for coronals, the tongue dorsum gesture for palato-velars.

Partisans of auditory theories have also searched possible invariant auditory cues that characterize the plosive place of articulation. The pioneer work by Delattre, Liberman, & Cooper (1955) on the “acoustic locus” actually served as a precursor for both auditory and motor proposals on invariance. In the framework of his Quantal Theory, Stevens proposed at the end of the 1970s that there might be a local spectral invariant for the plosive place of articulation, located around the position of the acoustic burst and independent of the speaker, the plosive manner of articulation and the context. Bilabial spectra would be “diffuse falling” (with energy all over the spectrum but more of it at low frequencies), alveolars would be

"diffuse raising" (idem but with more energy at high frequencies) and velars would be compact (with most of the energy packed into the medium) (Blumstein & Stevens, 1979; Stevens, 1980; Stevens & Blumstein, 1978). Following further proposals by Kewley-Port (1983), a progressive shift was made towards dynamic cues associating spectral values with the plosive and the next vowel. At the end of this process, the locus came back with the "locus equations" introduced by Sussman (Sussman, Fruchter, Hilbert, & Sirosh, 1998; Sussman, Hoemeke, & Ahmed, 1993; Sussman, McCaffrey, & Matthews, 1991) assuming relational invariance (correlations between $F2$ values for the plosive and the next vowel) as a correlate of the place of articulation. Importantly, acoustic characterization of the plosive place of articulation seems basically to rely on spectral data at two instants; plosive release and vowel climax.

Our proposal is different. In the PACT framework and in light of the perceptuo-motor developmental schedule described at the beginning of this paper, we presume that in a first stage, speech perception would benefit from rapidly maturing auditory processing that enables infants to categorize all CV sequences available in their environments. In this first stage, the motor system would not be mature and probably not even completely functionally related to the speech perception system according to Kuhl et al. (2014). Hence, the infants would not have at their disposal invariant cues for the plosive place of articulation. This question is debated, with negative results on plosive invariance before 6 months in Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Eimas, 1999; vs. data suggesting the possibility to discriminate /b/ from /d/ at 6 months of age, Hochmann & Papeo, 2014; and a discussion on possible confounding effects in Dole, Loevenbruck, Pascalis, Schwartz, & Vilain, 2015.

In a second stage, after 7 months there is progressive coupling and maturation of the speech motor system. Then, infants could discover that plosive-vowel sequences heard in the environment are produced by specific movements of the lips for bilabials, and the tongue apex

or dorsum for alveolars or velars. Hence, the content of the motor repertoire would enhance perceptual representations and allow invariance to emerge in a perceptuo-motor space.

For vowels, the situation is probably different. Indeed, auditory representations for oral vowels have been described in a number of studies, and oral vowels seem properly characterized in all their phonetic dimensions in a bundle of frequency parameters (e.g. ($F1$-F0) for height, ($F2$-F1) for place of articulation and $F'2$ for rounding; all values are in Barks: see Ménard, Schwartz, Boë, Kandel, & Vallée (2002). In contrast, the articulatory characterization of oral vowels is less straightforward (e.g., Boë, Perrier, & Bailly, 1992) and perturbation experiments suggest that invariants for vowels could be auditory rather than motor (Savariaux, Perrier, & Orliaguet, 1995; Savariaux, Perrier, Orliaguet, & Schwartz, 1999). It is more in terms of vowel reduction that articulatory dynamics could play a role, though the debate on this topic was vigorous in the 1980s and 1990s (e.g., Strange (1989) vs. Nearey (1989) or Perrier, Lœvenbruck, & Payan (1996) vs. Pitermann (2000)).

Therefore, our hypothesis is that the auditory and motor systems could be complementary in terms of the content of their representations for phonetic invariance, motor or gestural cues probably being crucial for the plosive place of articulation, while auditory parameters would be efficient for vowel characterization [(4)]. This is what we now propose to test with COSMO. For this aim, since natural articulatory data are sparse, particularly about perceptuo-motor development early in life, we will use synthetic data in the framework of the articulatory model of the vocal tract, *VLAM*.

VLAM and the generation of synthetic CV syllables

VLAM is a realist vocal tract model (Maeda, 1990) thanks to which seven articulatory parameters (*Jaw, Larynx, TongueBody, TongueDorsum, TongueApex, LipHeight, LipProtrusion*) have been derived from a guided principal component analysis of

cineradiographic images of the vocal tract. These allow the description of the jaw and larynx position, and of the tongue and lips shape. The parameters can be interpreted in terms of phonetic and muscular commands (Maeda & Honda, 1994). The areas of 28 sections of the vocal tract are estimated as linear combinations of these seven parameters, which then allows computation of the transfer function and formants (Badin & Fant, 1984) (see Figure 9).

FIGURE 9 ABOUT HERE

In short, *VLAM* is a geometric model enabling formants from articulatory parameters to be computed. This model has been evaluated over the last fifteen years in terms of its ability to generate vowels and plosive stimuli compatible with data from infants (Boë et al., 2013), children and adults (Laurent et al., 2013; Ménard et al., 2002; Ménard, Schwartz, & Aubin, 2008; Ménard, Schwartz, & Boë, 2004; Schwartz, Boë, Badin, & Sawallis, 2012b). It is also the articulatory synthesizer of the DIVA (Directions Into Velocities of Articulators) model of speech production (Guenther, 2006; Guenther, Hampson, & Johnson, 1998).

Here, VLAM is considered as a simplified implementation of the motor-to-auditory relationship in the human vocal tract [(5)]. It is used both to generate CV syllables thought to be produced by the Master Agent and as an external simulator of the Learning Agent' vocal tract so that it can learn from the perceptual consequences of the motor commands it is sending. VLAM also incorporates a model for vocal tract scaling associated with age, thanks to which the size increases with age in a nonlinear way compatible with experimental data (see Boë et al., 2013; Ménard et al., 2002, 2008). However, as in Part 2, we do not consider here vocal tract differences between the Learning Agent and the Master, supposing that if there were any, they could be solved by appropriate normalization processes (see Ménard et al., 2002).

Generation of oral vowels

Vowels are defined as articulatory configurations that are not too closed, so as not to generate noise in their acoustic output. This is characterized in VLAM by setting a constraint

on the constriction, which is the position of the section of the vocal tract with the smallest area. The constriction area for vowels is higher than a minimum value of 0.15 cm^2. In the present set of simulations, we only considered the three extreme oral vowels /i, a, u/, which provide the preferred choice in human languages (see Schwartz, Boë, Vallée & Abry, 1997).

Any speech sound should need all 7 VLAM parameters for its complete generation and characterization. However, we have attempted to keep the number of free parameters at the smallest possible value to minimize later computations. Hence, vowels are described here by three *VLAM* articulatory parameters (*TongueBody*, *TongueDorsum* and *LipHeight*), all other parameters being set to a neutral value (resting position). We define motor vowel prototypes for /a i u/, using average formant values for French vowels (Meunier, 2007) as targets and selecting values of the three *VLAM* parameters that best fit the acoustic target. For each category of vowel, we generated a set of articulatory configurations according to a Gaussian probability distribution centered on the prototype value.

Generation of stop consonants

Plosives are defined as articulatory configurations achieved just after a complete closure of the vocal tract, i.e. at the time of acoustic release, which typically generates an acoustic burst. In the present simulations we characterize plosives by the formants produced with a constriction close to, but still slightly higher than, zero, so as to be able to compute formants. We only considered the three extreme plosive places of articulation (labial, alveolar, velar) that provide the preferred choice in human languages (see Schwartz et al., 2012b). The unvoiced stop consonants /p, t, k/ corresponding to these places of articulation are more frequent in human language than their voiced counterparts /b, d, g/, but, in the rest of this paper, we keep the voiced set of consonants /b, d, g/, because voiced plosives provide the clearest formant trajectories and enable a better specification of formants at the beginning of the opening trajectory from the plosive to the next vowel.

We adopt the view proposed by Öhman (1966) that plosives are local perturbations (vocal tract closing gestures) of vowel configurations within CV syllables. Therefore, we synthesize plosives by closing the vocal tract from a vowel position, using *the VLAM Jaw* parameter combined with one other articulator, i.e. *Jaw* and *LipHeight* for /b/, *Jaw* and *TongueApex* for /d/, and *Jaw* and *TongueDorsum* for /g/. Hence, plosives are described by five parameters (*Jaw, TongueBody, TongueDorsum, TongueApex* and *LipHeight*). Furthermore, the perturbation gesture allowing a consonant to be produced from a vowel is characterized by two parameters: the variation (Delta) of *Jaw* and another one from among *LipHeight*, *TongueApex* or *TongueDorsum*. To obtain a consonant, both articulators should be combined, so that the vocal tract constriction area reaches a value between 0.05 and 0.15 cm². More specifically, the set of consonants that can be achieved from a vowel configuration of the vocal tract is the set of configurations obtained by 1) going through all possible discrete values of the parameter *Jaw* and 2) for each of these values selecting the value of the other articulator (*LipHeight*, *TongueApex* or *TongueDorsum*) such that when the perturbation is applied to the vowel the constriction area is the closest possible to 0.05 cm². The choice of modeling a consonant as a perturbation added to a vowel means that consonants and vowels are linked by maximal co-articulation.

Representation of CV sequences

Once stop consonants and vowels have been defined in terms of articulatory and acoustic parameters, the question is to define an adequate representation of the trajectory from C to V, characterizing the syllable in articulatory and acoustic terms. Since we showed in the previous section that the data converge towards a characterization based on plosive onset and vowel formants, plosive-vowel syllables are characterized as a pair of two articulatory states: one for the plosive and the other for the vowel, neglecting the geometry and temporal aspects of the trajectory linking these two states. Altogether, a CV sequence is associated in VLAM

with 8 articulatory parameters, 5 for the plosive and 3 for the vowel.

In the acoustic space, vowels are characterized by their first two formants $(F1, F2)$,which *VLAM* computes from the articulatory parameters in the open state. For plosives, where $F1$ is basically the same for all configurations (around 250 Hz), characterization is by $F2$ and $F3$, computed by *VLAM* in the closed state. A CV sequence is associated to 4 acoustic parameters, 2 for the plosive and 2 for the vowel.

FIGURE 10 ABOUT HERE

Figure 10 displays the acoustic properties of the vowels and plosives generated. The representation of vowels in the $(F1, F2)$ plane is classical, with /i, a, u/ at the corners of the vowel triangle (Figure 10, top). The representation of plosives in the $(F2, F3)$ plane is less common (Figure 10, bottom). It has been extensively discussed in Schwartz et al. (2012b). We observe that there is a trend toward lower $(F2, F3)$ values for /b/, higher values for /d/ and medium values for /g/. This recalls the "diffuse falling" vs. "diffuse raising" vs. "compact" contrasts proposed by Stevens and Blumstein (1978), but with considerable variations of the plosive recognition depending on the vowel context.

FIGURE 11 ABOUT HERE

We show in Figure 11 the relationship between the $F2$ values for plosives and vowels, providing a portrait that is globally coherent with the one reported by Sussman et al. (1998) for natural speech.

The two-state implementation of syllabic trajectories is highly simplified in relation to natural CV sequences, and a number of more elaborate CV co-articulation models have been suggested since the pioneer one proposed by Öhman (1966). However, here we merely aimed to generate syllables whose variability patterns were similar to the complexity of real speech signals. The syllable material displayed in Figure 10 provides an adequate compromise. It corresponds to complex variations in an 8-D articulatory space and resulting in variations in a

4-D acoustic space with co-articulation patterns that are globally coherent with those of natural syllables. We will next examine how the motor and auditory systems of the *COSMO* model extended to syllables can deal with this variability.

***COSMO-S*, an extension of the *COSMO* model to process plosive-vowel syllables**

FIGURE 12 ABOUT HERE

We have extended the *COSMO* model to CV syllable processing. The objects, O_S from the speaker's point of view, and O_L from a listener's perspective, refer to the syllables we consider: /ba/, /bi/, /bu/, /ga/, /gi/, /gu/, /da/, /di/, /du/. Since we model a syllable as a vowel state and a consonant state, the variable S separates into S_V and S_C, and the variable M into M_V and M_C. Apart from that, the *COSMO-S* model (see Figure 12, top) shares its global structure with COSMO as it is made of the same systems: (i) the auditory system associates sensory representations with the corresponding O_L syllable labels; (ii) the sensory-motor system associates motor and sensory representations; (iii) the motor system associates motor representations with O_S syllable labels.

These systems are linked by λ coherence variables, which are a mathematical tool used to force duplicate variables to have the same values at all times during probabilistic inference (Bessière et al., 2013; Gilet, Diard, & Bessière, 2011). This provides a mathematical implementation of a probabilistic switch, allowing the different parts of the model to be activated or deactivated during probabilistic inference, thus permitting constraints coming from the different sub-models to be integrated into the global model. Likewise, the specification of $C = True$ in an inference task allows the combination of motor and auditory cues.

The auditory system describes the knowledge the agent has of the link between O_L syllables and sensory variables: ${S'}_V$ ($F1$ and $F2$ for the vowel) and ${S'}_C$ ($F2$ and $F3$ for the

consonant). These are implemented as 4-D Gaussian probability distributions, the mean vectors and covariance matrices of which are estimated during the learning process (see below).

The sensory-motor system describes the knowledge the agent has of the motor-to-sensory mapping, i.e. of mapping between articulatory gestures M_V (vowel), M_C (consonant) and formant values S_V and S_C. Once again, mappings are described by Gaussian probability distributions, where mean vectors and covariance matrices are estimated during the learning process (see below). The term $P(M_C \mid M_V)$ encodes a support for consonants that can be achieved according to the perturbation hypothesis described in the section 'Generation of stop consonants'. More specifically, for each vowel motor gesture M_V, $P(M_C \mid M_V)$ defines a probability distribution that is a plateau in the 5-D articulatory space for consonants. It is uniform on the possible attainments of consonants obtained by the joint use of the parameter *Jaw* and another one (either *LipHeight* for /b/, *TongueApex* for /d/ or *TongueDorsum* for /g/), and it is null everywhere else for configurations that are not consonants (because the vocal tract is not closed enough) or for configurations that cannot be reached from the M_V vowel configuration considered. The term $P(M_C \mid M_V)$ implements a constraint coming from the physics of the Learning Agent's vocal tract (modeled by VLAM), which does not have to be estimated in the learning stage. This constraint is implemented using conditional probability tables, assigning a constant value to each achievable consonant gesture and zero probability otherwise.

The motor system describes a state of knowledge of the link between O_S syllable labels and articulatory gestures. The structure of the motor system implements a simplified co-articulation model based on Öhman's perturbation hypothesis (Öhman, 1966). This explicitly introduces a delta variable describing the perturbation superimposed on the vowel to obtain a plosive consonant. Furthermore, we assume in *COSMO-S* that the Learning Agent

would have at its disposal a set of primitive consonant gestures corresponding to the basic places of articulation for plosives: combined jaw and lips for bilabials, combined jaw and tongue apex for alveolars, and combined jaw and tongue dorsum for velars. The learning process would consist of discovering these basic primitive gestures through motor exploration, and identifying their correspondence with the CV sequences provided by the Master Agent. Therefore, while vowels in the motor repertoire are characterized by their articulatory configuration M'_V (*TongueBody*, *TongueDorsum* and *LipHeight* in *VLAM*), plosives are characterized by their primitive gesture G'_C, referring to the articulator used to make a plosive consonant in coordination with *Jaw* (*LipHeight* for /b/, *TongueDorsum* for /g/, and *TongueApex* for /d/. G'_C is thus a categorical variable, with three possible values). Variables M'_V and G'_C are taken to be independent. Hence $P(M'_V\ G'_C \mid O_S) = P(M'_V \mid O_S).P(G'_C \mid O_S)$. The motor configuration of the plosive in the framework of Öhman's perturbation theory is then defined by Δ'_{MC}, the variation of the articulators (the specific combination of *Jaw* and another specific articulator) necessary to achieve a consonant from M'_V. The motor command for the M'_C consonant is finally obtained by the equation $M'_C = M'_V + \Delta'_{MC}$. The term $P(\Delta'_{MC} \mid M'_V\ G'_C)$ describes how the consonant is produced, depending on the vowel and the specific consonant gesture. This shows explicitly that the consonant is conditioned by the vowel, which can be interpreted as anticipation. For instance, to produce the sound /ba/, the /a/ is anticipated when /b/ is performed, which amounts to having maximal co-articulation. $P(M'_V \mid O_S)$ and $P(\Delta'_{MC} \mid M'_V\ G'_C)$ are described by Gaussian distributions, where the mean vectors and covariance matrices are estimated during the learning process (see below). Finally, $P(G'_C \mid O_S)$ is implemented with a conditional probability table (histogram), the parameters of which are also identified during learning.

The COSMO-S model is thus defined by the joint probability distribution

decomposition shown in Figure 12 (bottom).

Similarly to the summary of Figure 2, the Bayesian inference within the *COSMO-S* model allows computing of conditional probability distributions. Purely motor, purely auditory and perceptuo-motor instances of the speech perception task are implemented. Because of the complexity of the COSMO-S model, we have not detailed the corresponding Bayesian inferences here. However, they can be interpreted exactly as previously: auditory perception is expressed as direct use of the link between auditory representations and the corresponding object labels, motor perception as the combination of the motor repertoire with an internal model allowing association of motor and sensory representations, and perceptuo-motor perception as the Bayesian fusion of the auditory and motor categorization processes. We will now describe how the Learning Agent acquires the different parts of the model.

Learning in *COSMO-S*

Some probability distributions of the model are not learned. Indeed, the prior $P(O_S^{Ag})$, $P(O_L^{Ag})$ and $P(M_V^{Ag})$ are set as uniform probability distributions. The biological constraints $P(M_C^{Ag} \mid M_V^{Ag})$ describe the consonants achievable from a given vowel, and are pre-computed in *VLAM*. Finally, probability distributions over coherence variables, $P(\lambda_{SV}^{Ag} \mid S_V^{Ag} S'^{Ag}_V)$, $P(\lambda_{SC}^{Ag} \mid S_C^{Ag} S'^{Ag}_C)$, $P(\lambda_{MV}^{Ag} \mid M_V^{Ag}\ M'^{Ag}_V)$, $P(\lambda_{MC}^{Ag} \mid M_C^{Ag}\ \Delta'^{Ag}_{MC}\ M'^{Ag}_V)$ and $P(C^{Ag} \mid O_S^{Ag}\ O_L^{Ag})$ are set as Dirac probability distributions, with $True$ value of a probability of 1 for a given relationship between the variables on the right hand side, respectively $S_V^{Ag} = S'^{Ag}_V$, $S_C^{Ag} = S'^{Ag}_C$, $M_V^{Ag} = M'^{Ag}_V$, $M_C^{Ag} = \Delta'^{Ag}_{MC} + M'^{Ag}_V$ and $O_S^{Ag} = O_L^{Ag}$.

The probability distributions that the Learning Agent apprehends in *COSMO-S* are the same as in the 1-D implementation studied in the previous section: the auditory categorization branch $P(O_L^{Ag} \mid S_V^{Ag}\ S_C^{Ag})$, the forward model implementing the motor-to-auditory

relationship $P(S_V^{Ag}\ S_C^{Ag} \mid M_V^{Ag}\ M_C^{Ag})$ and the motor repertoire $P(M_V^{Ag}\ M_C^{Ag} \mid O_S^{Ag})$. As previously, we learn the auditory and motor branches independently from each other, but with the same set of data. This allows a fair comparison between the two branches.

While the forward model and the motor repertoire were learned in 1-D, a two-stage process was implemented in *COSMO-S*. Indeed, considering the complexity of the motor-to-auditory relationship within a 12-D space (8 motor plus 4 auditory dimensions), it appeared easier to learn the forward model before the motor repertoire. This corresponds well to the developmental schedule presented previously (Kuhl, 2004), which led us to proceed in three consecutive steps:

L1. learning the auditory categories;

L2. learning motor-to-auditory mapping;

L3. learning the motor repertoire.

During these three learning phases, the Learning Agent interacts with a Master Agent to obtain syllable acoustic stimuli ($F2, F3$ for the plosive, $F1, F2$ for the vowel) taken from the data displayed in Figure 10 and, for steps L1 and L3, the corresponding syllable labels as well. Phases L2 and L3 are independent of phase L1; hence they will be evaluated separately in the following argument.

L1: Learning the auditory categories by association.

As in our previous experiments, the auditory system, linking auditory representations S'^{Ag}_V and S'^{Ag}_C and corresponding syllables O_L^{Ag}, is learned by association, through interactions with the Master Agent. More precisely, the term $P(S'^{Ag}_V\ S'^{Ag}_C \mid O_L^{Ag})$ consists of 9 auditory prototypes (one for each value of O_L^{Ag}) encoded as 4-D Gaussian probability distributions on the formant space ($F1_V$, $F2_V$, $F2_C$, $F3_C$), which the agent learns in a supervised manner from the Master Agent. This provides <formant values, syllable label> pairs. Auditory recognition

$P(O_L^{Ag} \mid S'^{Ag}_V S'^{Ag}_C)$ is then implemented by the Bayesian inversion of $P(S'^{Ag}_V\ S'^{Ag}_C \mid O_L^{Ag})$:

$$P(O_L^{Ag} \mid S'^{Ag}_V S'^{Ag}_C) = \frac{P(S'^{Ag}_V\ S'^{Ag}_C \mid O_L^{Ag})}{\sum_o P(S'^{Ag}_V\ S'^{Ag}_C \mid [O_L^{Ag}=o])}.$$

L2: Learning the motor-to-auditory mapping by accommodation.

Since we attempt to learn the sensory-motor system independently of the motor repertoire, learning is achieved by a variant of the learning by accommodation algorithm, in which the Learning Agent only obtains auditory input from the Master Agent, without object labels. Given a syllable acoustic target (s_V, s_C), and using its current state of knowledge as given by $P(S_V^{Ag}\ S_C^{Ag} \mid M_V^{Ag}\ M_C^{Ag})$, the Learning Agent carries out imitation tasks, by inferring a motor gesture (m_V, m_C) likely to reach the target. This gesture is obtained by randomly drawing a value (m_V, m_C) according to the inversion of the current forward model:

$$P\left(M_C^{Ag}\ M_V^{Ag} \mid [S_V^{Ag} = s_V][S_C^{Ag} = s_C]\right) \propto$$

$$P\left(M_V^{Ag}\right).P\left([S_V^{Ag} = s_V] \mid M_V^{Ag}\right).P\left(M_C^{Ag} \mid M_V^{Ag}\right).P\left([S_C^{Ag} = s_C] \mid M_C^{Ag}\right).$$

The gesture (m_V, m_C) is sent to *VLAM*, which plays the role of an external vocal tract simulator. *VLAM* outputs the formants (s_V^*, s_C^*) corresponding to the motor command (m_V, m_C), and the Learning Agent updates the knowledge stored in its internal models. It observes that the chosen motor commands produce a given set of formants. This knowledge is stored in the probability distributions $P\left(S_V^{Ag} \mid [M_V^{Ag} = m_V]\right)$ and $P\left(S_C^{Ag} \mid [M_C^{Ag} = m_C]\right)$, which are Gaussian probability distributions evolving as their parameters become updated through the learning process.

The syllable targets provided by the Master Agent to the Learning Agent are taken from the data presented in Figure 10. Since the Learning Agent initially has no knowledge available, it selects motor gestures randomly. New observations lead to improving the quality of the internal model of the motor-to-auditory transformation, which in turn improves the

motor inversion that relies on this internal model. This means that the computed probability distribution $P\left(M_C^{Ag}\ M_V^{Ag} \mid \left[S_V^{Ag} = s_V\right]\left[S_C^{Ag} = s_C\right]\right)$ driving the choice of motor gestures and allowing imitation of auditory inputs becomes more and more accurate. Thus, the agent becomes better and better at reaching its targets. All along the exploration process, the learning algorithm remains driven by the targets provided by the Master, rather than by an exhaustive sampling of the motor space as in other systems (e.g., Bailly, 1997; Guenther, 2006).

L3: Learning the motor repertoire by imitation.

The motor system is learned in a supervised way, in that syllable labels are given to the Learning Agent along with the corresponding stimuli. But while in other research the articulatory data are provided (Castellini et al., 2011; Canevari et al., 2013), here the Learning Agent is only given labeled acoustic data. We use the same <formant values, syllable label> pairs that served to learn auditory categorization in step L1, and we use the internal model of the motor-to-auditory transformation learned in step L2 to retrieve motor information. Given an acoustic target (s_V, s_C) and the corresponding syllable label o_S, the Learning Agent infers a motor gesture allowing the target to be reached by inversing the motor-to-auditory mapping and by using its present state of knowledge of the correspondence between syllables and motor gestures. This is done by randomly drawing $(m'_V, g'_C, \delta'_{MC})$ values according to the following probability distribution:

$$P\left(\begin{matrix}\left[M'^{Ag}_V = m'_V\right]\left[G'^{Ag}_C = g'_C\right] \\ \left[\Delta'^{Ag}_{MC} = \delta'_{MC}\right]\end{matrix}\middle|\begin{matrix}\left[S_V^{Ag} = s_V\right]\left[S_C^{Ag} = s_C\right]\left[O_S^{Ag} = o_S\right] \\ \left[\lambda_{MV}^{Ag} = 1\right]\left[\lambda_{MC}^{Ag} = 1\right]\end{matrix}\right)$$

$$\propto \begin{pmatrix} P\left(\left[M_V^{Ag} = m'_V\right]\right) P\left(\left[S_V^{Ag} = s_V\right] \mid \left[M_V^{Ag} = m'_V\right]\right) \\ P\left(\left[M_C^{Ag} = m'_V + \delta'_{MC}\right] \mid \left[M_V^{Ag} = m'_V\right]\right) P\left(\left[S_C^{Ag} = s_C\right] \mid \left[M_C^{Ag} = m'_V + \delta'_{MC}\right]\right) \\ P\left(\left[M'^{Ag}_V = m'_V\right]\left[G'^{Ag}_C = g'_C\right]\left[\Delta'^{Ag}_{MC} = \delta'_{MC}\right] \mid \left[O_S^{Ag} = o_S\right]\right) \end{pmatrix}$$

The correspondence between the chosen motor gesture $(m'_V, g'_C, \delta'_{MC})$ and the syllable label o_S is then used to update parameters of the following probability elements: the Gaussian probability distribution $P\left(M'^{Ag}_V \mid O^{Ag}_S\right)$, the histogram $P\left(G'^{Ag}_C \mid O^{Ag}_S\right)$ and the Gaussian probability distribution $P\left(\Delta'^{Ag}_{MC} \mid M'^{Ag}_V \; G'^{Ag}_C\right)$.

Simulation results

Confirming the "auditory-narrowband vs. motor-wideband" portrait in *COSMO-S*.

The aim of the simulations described in this section is to verify how the main principles we extracted from the results of the experiments carried out in the 1-D case are generalized to the more realistic case of syllable processing. We ran a single learning simulation with 4,000,000 <formant values, syllable label> pairs. The first 3,000,000 were used during the L2 learning phase, with another 1,000,000 during L3 and the full set of the same 4,000,000 values were also used for the L1 learning phase. These were resampled from the data presented in Figure 10 and provided to the Learning Agent by the Master Agent. Since the L1 learning phase of the sensory model is independent from the L2 and L3 learning phases of the perceptuo-motor and motor models and for future comparisons of the auditory and motor models of perception to be fair, the same data is used as input to learn the components involved in motor and auditory perception.

FIGURE 13 ABOUT HERE

Firstly, in Figure 13 we compare the evolution through learning of the entropy $H\left(P\left(S^{Ag} \mid O^{Ag}_L\right)\right)$ of the auditory model, with the evolution of the entropy $H\left(P\left(S^{Ag} \mid O^{Ag}_S\right)\right)$ of the motor model. To this end, as in Figure 5 we used the entropy $H\left(P\left(S^{Ag} \mid O^{Master}_S\right)\right)$ of the probability distribution over the stimuli produced by the Master Agent, a constant over the learning process, as a reference. As in the 1-D case, we further average these entropies over

objects; since we now have 9 possible objects, we consider $\frac{1}{9}\sum_{O_L^{Ag}} H\left(P\left(S^{Ag} \mid O_L^{Ag}\right)\right)$ for the auditory model, $\frac{1}{9}\sum_{O_S^{Ag}} H\left(P\left(S^{Ag} \mid O_S^{Ag}\right)\right)$ for the motor model, and $\frac{1}{9}\sum_{O_S^{Master}} H\left(P\left(S^{Ag} \mid O_S^{Master}\right)\right)$ as the Master's reference.

We observe that, as with the 1-D model, the entropy of the auditory model converges quickly to a level close to the entropy of the stimuli produced by the Master, whereas the entropy of the motor model converges more slowly(6).

FIGURE 14 ABOUT HERE

To better display how exploration and learning proceed in *COSMO-S*, we show in Figure 14 the motor gestures m'_V selected by the Learning Agent to attempt to reproduce the auditory targets provided by the Master Agent at five stages in the learning process: before any learning took place, during the L2 learning phase, between L2 and L3, during L3, and at the end of the learning process. We observe that learning enables progressive focusing of the vowel articulatory gestures around given areas in the articulatory space, corresponding to adequate commands for the three vowels /a, i, u/, but the size of the possible motor configurations remains wide.

FIGURE 15 ABOUT HERE

Comparatively, we display in Figure 15 the actual productions of the Learning Agent in acoustic space at the same five steps in the learning process: the size of the available auditory space is more reduced around the three vowels /a i u/ provided by the Master Agent at the end of the learning process.

FIGURE 16 ABOUT HERE

To evaluate the global categorization ability of the auditory, motor and perceptuo-motor branches in *COSMO-S* at the end of the learning process with 4,000,000 iterations, we exploited the same methodology as with the 1-D implementation of *COSMO*. We took as

input the formant values ($F2, F3$ for the consonant, $F1, F2$ for the vowel) produced by the Master Agent (and displayed in Figure 10). We added a given level of noise by adding a Gaussian perturbation to each formant value, with a given variance indexed by the noise level. We present these stimuli to the auditory, motor, or perceptuo-motor classifier defined in *COSMO-S* according to the equations derived from Figure 2. In practice, we used exact inferences for evaluation: stimuli were not sampled from the Master Agent and then decoded, rather the whole probability distribution of stimuli was used to directly compute the resulting confusion matrices for each classifier. The original object O_S^{Master} was compared with the decoded object O_L^{Agent} or O_S^{Agent} depending on the model considered. Average diagonal values of the confusion matrices provide recognition scores that are displayed in Figure 16.

The pattern of results is similar to that obtained with 1-D simulations. For clean stimuli (Figure 16, noise = 0), the auditory model is more accurate than the motor one. The difference is small, considering the large difference in entropies at the end of the learning process (see Figure 13), but this is because the distributions to categorize are well separated in these simulations. The difference would be larger in less clean learning configurations. When noise is added, the motor system performance decreases less rapidly than the auditory one, and it becomes more accurate for noise levels greater than 0.5. The perceptuo-motor model capitalizes on the fusion of the two branches to provide better scores than the separate auditory and motor models, at all noise levels.

Assessing auditory vs. motor invariance for the place of articulation of vowels and plosives in *COSMO-S.*

We now explore our second proposal about auditory-motor complementarity, assessing how phonemic invariance could be represented in the auditory or motor branches in *COSMO-S*.

The situation for vowels has already been presented in Figures 14 and 15. It is in

agreement with our predictions; while the acoustic characterization of vowels is rather straightforward (see the final learning stage in Figure 15), the distribution of motor variables is more disordered (see the final learning stage in Figure 14).

FIGURE 17 ABOUT HERE

For plosives, the acoustic configuration is more complicated than for vowels. Indeed, Figure 10 shows how intricate the formant configurations are for each plosive, due to vowel co-articulation. This is where the motor system could play a crucial role. Indeed, in Figure 17 we display the evolution of the motor variable $P\left(G'^{Ag}_{C} \mid O^{Ag}_{S}\right)$ distribution for the 9 objects O^{Ag}_{S}. Each subplot displays the evolution of the probabilities of the three possible gestures (*LipHeight*, *TongueDorsum* and *TongueApex*) for each object at each learning stage. It appears that for 8 cases out of 9, the identification of the correct gesture has been successful: *LipHeight* for /ba, bi, bu/, *TongueDorsum* for /ga, gi, gu/ and *TongueApex* for /da, di, du/. This means that the Learning Agent has selected a gesture compatible with that performed by the Master Agent, and that motor invariance is within reach through the existence of the G'^{Ag}_{C} parameter in the motor repertoire.

However, there is one case where adequate identification has partly failed: for the object /gi/, the probability of the *TongueApex* gesture remains high, even though this is not the adequate gesture for the velar in /gi/. The reason is clear: looking at Figure 10, it can be seen that the acoustic regions for /di/ and for /gi/ are partially superimposed. This is probably due to an acoustic description of the plosives that is too simplified (e.g. lacking higher formants, burst fine characteristics or spectral dynamics, which could all play a part in improving the /di/-/gi/ contrast). However, even if there was an articulatory ambiguity, we may suppose that hyper-articulation by the Master Agent could guide the Learning Agent to solve the problem. In a further simulation, we implemented this process by having a Master Agent discarding productions before reaching an acoustic zone where both /di/ and /gi/ could be produced. With

such a dataset for /gi/ production, the recovery of adequate gestures is perfect, as displayed in Figure 18.

FIGURE 18 ABOUT HERE

Concluding Part 3: Summary and predictions

The simulations with COSMO-S enable a significant gap in complexity to be crossed and the possibility of implementing and testing the *COSMO* model in a high dimension (8 articulatory + 4 acoustic parameters) could be assessed. These provide two major results.

Firstly, we confirmed the auditory-narrow motor-wide portrait introduced in Part 2. Once again we obtain both quicker and more efficient learning of acoustic stimuli in the sensory compared with the motor pathway (Figures 13-15), resulting in better auditory recognition scores than motor ones without noise, but also a superiority of the motor decoding process in noisy conditions (Figure 16). In consequence, altogether the perceptuo-motor model performs better than both the auditory and the motor pathways whatever the noise level (Figure 16). This was expected, given that the auditory-narrow motor-wide hypothesis is considered to be generic and independent of the underlying specific implementation. However, it is of interest to confirm that it is displayed in a much more complex and realistic sensory-motor environment in COSMO-S compared with the 1-D simulations of Part 2. Interestingly, the motor-to-sensory transformation in COSMO-S, associated with VLAM, is no longer monotonous. Altogether, and unsurprisingly, the increase in complexity even results in a much larger difference in learning speed and efficiency between auditory and motor inference in COSMO-S compared with 1-D simulations (compare Figures 5 and 13).

Secondly, the analysis of the information content of auditory vs. motor representations lets a natural complementarity appear, with plosives on one side, difficult to characterize in the auditory space, but clearly associated with a motor or gestural invariant in the motor

repertoire, and vowels on the other side, for which the auditory characterization is more efficient than the motor one.

The potential limitations of this study are related to the nature of the hypotheses we introduced to make the implementation of tractable simulations possible. Firstly, we had to base our work on artificial synthetic CV sequences. Indeed, *COSMO* requires a sensory-motor model able to process stimuli characterized in the motor (articulatory) and sensory (acoustic) spaces. No currently available speech production model can produce completely realistic speech stimuli. Thus we needed to restrict our simulations to synthetic stimuli generated by the model we had at our disposal, i.e. *VLAM*. However, the realism of formant data in our simulations, and the relative complexity of the material that we provided for processing in *COSMO-S*, make us confident that real speech is not out of reach of *COSMO* development.

A second hypothesis in *COSMO-S* is the assumption that the Learning Agent has at its disposal a set of primitive coordination gestures, i.e. *Jaw/Lips*, *Jaw/Tongue Apex* or *Jaw/Tongue Dorsum*, corresponding respectively to labial, alveolar and velar places of articulation. This hypothesis deserves further comment. It is consistent with data on infant imitation showing that infants have at their disposal basic facial gestures that they can identify from birth on the face of their communicating partner (Meltzoff & Moore, 1977). The Articulatory Organ Hypothesis developed in the Haskins Labs at the beginning of the 2000s (see e.g., Best & McRoberts, 2003; Goldstein & Fowler, 2003) exploits precisely this kind of assumption to describe perception and control in the framework of Articulatory Phonology. It supposes that infants are able to detect in a speech signal the primary articulatory organ that produced it. Current simulations provide this hypothesis with some computational basis.

These two results, “auditory-narrow, motor-wide” and “auditory-vowel, motor-plosive” provide two major sets of experimental predictions.

Prediction 3 - early speech perception should be mostly auditory before the onset of babbling, then become progressively perceptuo-motor

The PACT proposal is that speech perception should make use of only the auditory pathway in the first months of age, then progressively capitalize on feedback from the motor system when it is mature and mainly since babbling onset around 7 months. This appears to be reinforced in the context of the auditory-narrow, motor-wide hypothesis. The learning pattern in Figure 13 strongly confirms the view that auditory perception should be mature long before motor information could be used for phonetic decoding. Considering the potential role of the motor system for perception in noisy conditions, it could be suggested that as this is still immature at the first months of age, it should not intervene specifically in adverse conditions, before some significant degree of sensory-motor development. Basically, it is after babbling onset that infants can obtain a useful amount of information on the motor inference branch.

This is exactly what was described in a recent MEG (Magnetoencephalography) study on infants' brain responses to native vs. non-native stimuli at two developmental stages in the first year of age (Kuhl et al., 2014). Indeed, the data in this study showed that infants at 7 months of age do not display a significantly different involvement of the motor regions (including Broca's area and the cerebellum) for native vs. non-native speech. In contrast, at 11-12 months of age, i.e. after a significant amount of perceptuo-motor learning has occurred, following babbling onset at around 7 months, there is more involvement of motor regions for non-native compared with native stimuli.

Prediction 4 - plosive place of articulation invariance should require motor knowledge

The last prediction is related to the second result obtained with COSMO-S about the

"auditory-vowel, motor-plosive" hypothesis. The corresponding results in COSMO-S suggest that the identification of invariant cues for the plosive place of articulation should strongly depend on the acquisition of motor representations associated with sensory input in the motor inference process.

As mentioned in the Introduction, a number of experimental data do indeed suggest that infants cannot detect the plosive place of articulation invariance before 6 months of age. A recent study by Hochmann & Papeo (2014) exploiting a novel methodology based on pupillometry provided a hint that the "b" vs. "d" contrast could be displayed independently in vowel context at 6 months. However, auditory and visual information could be at the basis of this result (Dole et al., 2015). Importantly, another recent study in our group, exploiting an inter-sensory matching procedure, provided different results compatible with the present prediction. This procedure provided no evidence for articulation plosive identification independent of vowel context at 6 months of age, but some such evidence was seen at 9 months. Importantly, infants' perceptual abilities appeared to be related to their motor abilities in babbling (Dole, Loevenbruck, Pascalis, Schwartz, & Vilain, 2016).

The present prediction should not be generalized to the proposal that there would be no involvement at all of the motor system in speech perception before babbling onset. Indeed, Bruderer, Danielson, Kandhadai & Werker (2015) demonstrated that teething displays used to control infants' tongues in their mouths may interfere with the perception of non-native stimuli related to the corresponding induced tongue shapes for 6 month old subjects. The important point of our prediction is that the plosive place of articulation requires learning the sensory-to-motor correspondence in complete CV sequences that are out of reach before the onset of babbling.

Finally, it is of interest that a recent analysis of fMRI responses to CV syllables using multivariate decoding shows that plosive place of articulation is indeed specifically found to

be represented in regions of the brain associated with speech production, including the posterior ventral frontal cortex, the basal ganglia, and the cerebellum (Correia, Jansma & Bonte, 2015).

Part 4 - Three challenges for a perceptuo-motor theory of speech perception

At the end of this research, we have at our disposal the first Bayesian Perceptuo-Motor model of speech perception, COSMO, together with various implementations (from 1-D to COSMO-S). Furthermore, we have two major results about "non-perfect" learning conditions, enabling to depart from the indistinguishability theorem: the "auditory-narrow, motor-wide" and "auditory-vowel, motor-plosive" properties. This model opens a number of perspectives for future developments. We will discuss three major directions for research in the field of perceptuo-motor interactions involving potential developments in COSMO.

Challenge 1: Perceptuo-motor complementarity and Perceptuo-motor fusion

The Introduction showed how publications in the field shifted from almost purely functional arguments about the auditory vs. motor controversy, somewhat lacking of experimental data, to convincing experimental neurocognitive data supporting the role of the motor system, but somewhat lacking of functionalist views about why the motor system could be useful at all. The present study attempted to provide such functionalist arguments. Future studies should attempt to provide more data about when, how and why the motor system could enhance auditory perception. Furthermore, a perceptuo-motor theory of speech perception requires a fusion process enabling efficient combination of auditory (if not visual or somatosensory) and motor information for speech decoding.

Interestingly, audiovisual speech perception research has asked more or less the same questions for about the last forty years. It was shown how auditory and visual inputs could be complementary to a certain extent (e.g. Summerfield, 1987; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998). Audio-visual fusion led to many theoretical and methodological developments, proposing that it could be optimal in the Bayesian sense (see

Massaro and the Fuzzy-Logical Model of Perception, 1987, 1998; in relation with Ernst & Banks, 2002), and that within fusion each sensory modality could possibly be weighted according to its reliability, depending on context, language, subjects, etc. (Schwartz, 2010).

Perceptuo-motor complementarity and fusion should thus be set at a high position in the research agenda on perceptuo-motor speech perception, just as they were in past research on audiovisual speech perception.

Challenge 2 – Integrating speech perception and speech production in a common framework

Accumulating evidence for the role of motor knowledge in speech perception may be combined with accumulating evidence for the role of perceptual representations and processes in speech motor control (see reviews in Guenther, Hampson & Johnson, 1998; Perrier, 2005). Importantly, the current perceptuo-motor model of speech perception capitalizes on a set of computational bricks traditionally involved in speech production models. Thus, sensory and motor representations are associated thanks to internal forward or inverse models (e.g. Guenther, Ghosh & Tourville, 2006; Houde & Nagarajan, 2011; Hickok, 2012; Patri, Diard & Perrier, 2015).

This suggests that it could be possible to develop an integrated framework associating speech perception and speech production models within the same theoretical architecture. This is one aim of the COSMO architecture (see Moulin-Frier et al., 2012, 2015). Interestingly, the same objective has been introduced in recent phonological models (see e.g. Boersma, 2011, Boersma & Hamman, 2008).

Challenge 3 – From computational architecture to neurocognitive implementation

It is widely acknowledged, at least since Marr (1982), that cognitive systems can be

analyzed at different levels, three in Marr's proposal: computational, algorithmic, and representational and implementation levels. These levels are independent to a certain extent, but the computational and algorithmic architectures may shed light on the way neurocognitive implementation could be realized. Conversely, neurocognitive constraints could suggest some proposals for algorithmic considerations.

At this stage, we did not elaborate in any way the possible neurocognitive means by which the various components in COSMO could be implemented in the human brain. This is not to say that such an enterprise, relating computation and implementation levels, is out of reach, as was clearly displayed by the authors of the DIVA model of speech production (Guenther et al., 2006). Considering the increasing amount of details provided by neuroscience about the neural coding of speech perception and production in the auditory and motor cortex (see e.g. Bouchard, Mesgarani, Johnson & Chang, 2013; Cheung, Hamiton, Johnson, & Chang, 2016; Formisano, De, Bonte, & Goebel, 2008; Pasley et al., 2012), it is now a challenging but intriguing and probably necessary enterprise to attempt to elaborate further the possible relationships between computational models such as COSMO and neural responses in a number of experimental tasks.

Conclusion

This paper develops an original perspective in the debate between auditory and motor theories of speech perception. Research in the cognitive neurosciences led to the now well-accepted views that (i) motor areas are activated during speech perception, and (ii) motor knowledge seems to play a certain role in speech perceptual processing in the human brain. From these points of view, we attempted to evaluate the precise functional role of motor knowledge. In the framework of PACT, a perceptuo-motor theory of speech perception, we explored these questions in computational terms, thanks to COSMO, the first Bayesian

perceptuo-motor model of speech communication.

We showed here for the first time that, in conditions that are perfect in a certain sense, the information content of the auditory and motor branches of a perceptuo-motor speech processing system are exactly the same. We introduced realistic learning conditions and showed that they let a natural complementarity emerge between a “narrow-band” auditory system that is more efficient in good communication conditions, and a “wide-band” motor system that is more efficient in adverse conditions. Our simulations also suggest that invariants providing the phonetic characterization of phonological units could be perceptuo-motor rather than auditory or motor, and show how this could be achieved, with auditory cues for vowels and motor cues for the plosive place of articulation.

COSMO simulations lead to a number of experimental predictions. Some of these are already being tested, with data in agreement with predictions. Others require more experimental efforts. COSMO also opens a number of perspectives in domains such as: the fusion of perceptual and motor inference in phonetic decoding; the co-development of computational models of speech production and speech perception; the possibility to apply COSMO simulations to a number of neurocognitive data on the coding and processing of speech in the human brain.

This research has placed computational simulations at the heart of the debate about the role of perceptual and motor knowledge in the speech perception process. Considering the rapidly increasing amount of experimental evidence and data available about the perceptuo-motor relationship in speech communication, it seems that mathematical models can be of great help in clarifying arguments, precising mechanisms and suggesting new predictions and experimental paradigms. Perceptuo-motor complementarity, invariance, fusion and development are crucial steps in the agenda of future research into the cognitive bases of speech communication. The first pieces in the elaboration of the *COSMO* model described

and discussed in the present paper provide convincing elements for pursuing this direction.

1481 References

1482 Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*,
1483 93, 154–179. http://dx.doi.org/10.1037/0033-295X.93.2.154

1484 Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition.
1485 *Psychological Review*, 95, 124–150. http://dx.doi.org/10.1037/0033-295X.95.1.124

1486 d'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). The role of the motor system in
1487 discriminating normal and degraded speech sounds. *Cortex, 48*, 882–887.
1488 http://dx.doi.org/10.1016/j.cortex.2011.05.017

1489 d'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The
1490 motor somatotopy of speech perception. *Current Biology, 19*, 381–385.
1491 http://dx.doi.org/10.1016/j.cub.2009.01.017

1492 Badin, P. & Fant, G. (1984). Notes on Vocal Tract Computation. In *Quarterly Progress and Status*
1493 *Report*, Dept. for Speech, Music and Hearing, KTH, Stockholm (pp. 53–108).

1494 Bailly, G. (1997). Learning to speak. Sensori-motor control of speech movements. *Speech*
1495 *Communication, 22*, 251–267. http://dx.doi.org/10.1016/S0167-6393(97)00025-3

1496 Barnaud, M.L., Schwartz, J.L., Diard, J., & Bessière, P. (2016). Sensorimotor learning in a Bayesian
1497 computational model of speech communication. *6th International Conference on Development*
1498 *and Learning and on Epigenetic Robotics.*

1499 Basirat, A., Schwartz, J.-L., & Sato, M. (2012). Perceptuo-motor interactions in the perceptual
1500 organization of speech: Evidence from the verbal transformation effect. *Philosophical*
1501 *Transactions of the Royal Society B: Biological Sciences, 367*, 965–976.
1502 http://dx.doi.org/10.1098/rstb.2011.0374

1503 Bertoncini, J., Bijeljac-Babic, Blumstein, S., & Mehler, J. (1987). Discrimination in neonates of very
1504 short CVs. *Journal of the Acoustical Society of America, 82,* 31–37.
1505 http://dx.doi.org/10.1121/1.395570

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General, 117*, 21–33. http://dx.doi.org/10.1037/0096-3445.117.1.21

Bessière, P., Laugier, C., & Siegwart, R. (2008). *Probabilistic reasoning and decision making in sensory-motor systems*. Springer Tracts in Advanced Robotics. Berlin: Springer-Verlag.

Bessière, P., Mazer, E., Ahuactzin-Larios, J.-M., & Mekhnacha, K. (2013). *Bayesian Programming*. CRC Press. Boca Raton, FL.

Best, C.C., & McRoberts, G.W. (2003). Infant Perception of Non-Native Consonant Contrasts that Adults Assimilate in Different Ways. *Language and Speech, 46*, 183–216. http://dx.doi.org/doi:10.1177/00238309030460020701

Bever, T.G., & Poeppel, D. (2010). Analysis by Synthesis: A (Re-)Emerging Program of Research for Language and Vision. *Biolinguistics, 4.2-3,* 174–200. http://www.biolinguistics.eu

Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A., & Ward, B.D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience, 7*, 295–301. http://dx.doi.org/doi:10.1038/nn1198

Blumstein, S.E., & Stevens, K.N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66*, 1001–1017. http://dx.doi.org/10.1121/1.383319

Boë, L.-J., Badin, P., Ménard, L., Captier, G., Davis, B., MacNeilage, P., Sawallis, T., & Schwartz, J.-L. (2013). Anatomy and control of the developing human vocal tract: A response to Lieberman. *Journal of Phonetics, 41*, 379–392.

Boë , L.-J., Perrier, P., & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics, 20*, 27–38.

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer

model. *Acoustics Research Letters Online*, *4,* 129–134. http://dx.doi.org/10.1121/1.1613311

Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In Anton Benz & Jason Mattausch (Eds.) *Bidirectional Optimality Theory*, pp. 33–72. Amsterdam: John Benjamins.

Boersma, P., & Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint grammar. *Phonology, 25*, 217–270. http://dx.doi.org/10.1017/S0952675708001474

Bouchard, K.E., Mesgarani, N., Johnson, K., & Chang E.F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature, 495*, 327–332. http://dx.doi.org/10.1038/nature11911

de Boysson-Bardies, B. (1993). Ontogeny of language-specific syllabic production. In B. de Boysson-Bardies & S. de Schoen & P. Jusczyk & P. F. MacNeilage & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 353-363). Dordrecht: Kluwer Academic Publishers.

de Boysson-Bardies, B., Halle, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language, 16*, 1-17.

de Boysson-Bardies, B., Sagart, L., & Durant, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language, 11*, 1-15.

Bruderer, A.G., Danielson, D.K., Kandhadai, P., & Werker, J.F. (2015). Sensorimotor influence on speech perception in infancy. *Proc Natl Acad Sc, 112,* 13531-13536. http://dx.doi.org/10.1073/pnas.1508631112

Callan, D.E., Callan, A.M., & Jones J.A. (2014). Speech motor brain regions are differentially recruited during perception of native and foreign-accented phonemes for first and second language listeners. *Frontiers in Neuroscience*, 03 September 2014, http://dx.doi.org/10.3389/fnins.2014.00275

Callan, D.E., Jones, J.A., Callan, A.M., & Akahane-Yamada, R. (2004). Phonetic perceptual

identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *NeuroImage, 22*, 1182–1194, http://dx.doi.org/doi:10.1016/j.neuroimage.2004.03.006

Canevari, C., Badino, L., D'Ausilio, A., Fadiga, L., & Metta, G. (2013). Modeling speech imitation and ecological learning of auditory-motor maps. *Frontiers in Psychology, 4*, 364, http://dx.doi.org/10.3389/fpsyg.2013.00364

Castellini, C., Badino, L., Metta, G., Sandini, G., Tavella, M., Grimaldi, M., & Fadiga, L. (2011). The use of phonetic motor invariants can improve automatic phoneme discrimination. *PLoS ONE*, *6*, e24055. http://dx.doi.org/10.1371/journal.pone.0024055

Cheung, C., Hamiton, L.S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *eLife*, 5:e12577. DOI: 10.7554/eLife.12577

Clayards, M., Aslin, R. N., Tanenhaus, M. K., & Jacobs, R. A. (2007). Within category phonetic variability affects perceptual uncertainty. In *Proc. 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, pages 701–704.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108, 804–809. http://dx.doi.org/10.1016/j.cognition.2008.04.004

Correia, J.L., Jansma, B.M.B., & Bonte, M. (2015). Decoding articulatory features from fMRI responses in dorsal speech regions. *The Journal of Neuroscience, 35*, 15015–15025. http://dx.doi.org/10.1523/JNEUROSCI.0977-15.2015

Davis, B.L., MacNeilage, P., & Matyear, C.L. (2002). Acquisition of Serial Complexity in Speech Production: A Comparison of Phonetic and Phonological Approaches to First Word Production. *Phonetica, 59*, 75–107. http://dx.doi.org/10.1159/000066065

Dayan, P., & Abbott, L. (2001). *Theoretical Neuroscience*. The MIT Press, Cambridge, MA.

DeCasper, A.J., & Fifer, W.P. (1980). Of human bonding: newborns prefer their mother's voice. *Science, 208*, 1174-1176.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional Neuroimaging of Speech Perception in Infants. *Science*, *298*, 2013–2015. http://dx.doi.org/10.1126/science.1077066

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., & Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Sciences, 103*, 14240–14245, http://dx.doi.org/10.1073/pnas.0606302103

Delattre, P. C., Liberman, A. M. & Cooper, F. S. (1955) Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27*, 769–73. http://dx.doi.org/10.1121/1.1908024

Deng, L. (1999). Computational models for auditory speech processing,. In K. Ponting (Ed.) *Computational Models for Speech Pattern Processing*, pp. 67-77. NewYork: Springer-Verlag, NATO ASI.

Deng, L. & Ma, J. (2000). Spontaneous speech recognition using a statistical coarticulatory model for the vocal- tract-resonance dynamics. *The Journal of the Acoustical Society of America*, *108*, 3036–3048. http://dx.doi.org/10.1121/1.1315288

Deng, L., Ramsay, G., & Sun, D. (1997). Production models as a structural basis for automatic speech recognition. *Speech Communication, 22*, 93–111. http://dx.doi.org/10.1016/S0167-6393(97)00018-6

Diehl, R., Lotto, A., & Holt, L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149–179. http://dx.doi.org/10.1146/annurev.psych.55.090902.142028

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science*, 37, 344–377.

http://dx.doi.org/10.1111/cogs.12008

Dole, M., Loevenbruck, H., Pascalis, O., Schwartz, J.-L., & Vilain, A. (2015). Perceptual abilities in relation with motor development in the first year of life. *WILD 2015 International Conference*, Stockholm.

Dole, M., Loevenbruck, H., Pascalis, O., Schwartz, J.L, Vilain, A. (2016). Phoneme categorization depends on production abilities during the first year of life. XX ICIS 2016 - The International Congress on Infant Studies.

Eimas, P. D. (1999). Segmental and syllabic representations in the perception of speech by young infants. *The Journal of the Acoustical Society of America, 105*,1901–1911. http://dx.doi.org/10.1121/1.426726

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*, 303–306. http://dx.doi.org/10.1126/science.171.3968.303

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429–433. http://dx.doi.org/10.1038/415429a

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience, 15*, 399–402. http://dx.doi.org/10.1046/j.0953-816x.2001.01874.x

Feldman, N. H., & Griffiths, T. L. (2007). A rational account of the perceptual magnet effect. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pp. 257–262.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009a). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752–782. http://dx.doi.org/10.1037/a0017196

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009b). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pp. 2208–2213.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*, 751–778. http://dx.doi.org/10.1037/a0034245

Formisano, E., De, M.F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? brain-based decoding of human voice and speech. *Science, 322*, 970–973. http://dx.doi.org/10.1126/science.1164318

Fowler, C., & Dekle, D. (1991). Listening with eye and hand: crossmodal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 816–828. http://dx.doi.org/doi: 10.1037/0096-1523.17.3.816

Frankel, J., Richmond, K., King, S., et Taylor, P. (2000). An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. In *Proc. Sixth International Conference on Spoken Language Processing*, Vol. 4. International Speech Communication Association.

Galantucci, B., Fowler, C.A., & Turvey, M.T. (2006). The motor theory of speech perception reviewed. Psychonomic Bulletin & Review, 13, 361–377. http://dx.doi.org/10.3758/BF03193857

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613–656.

Gilet, E., Diard, J., & Bessière, P. (2011). Bayesian action–perception computational model: interaction of production and recognition of cursive letters. *PLoS ONE*, *6*, e20387. http://dx.doi.org/10.1371/journal.pone.0020387

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 159–207). Berlin: Mouton de Gruyter.

Grabski, K., Tremblay, P., Gracco, V.L., Girin, L., & Sato, M. (2013). A mediating role of the

auditory dorsal pathway in selective adaptation to speech: A state-dependent transcranial magnetic stimulation study. *Brain Research, 1515*, 55–65. http://dx.doi.org/10.1016/j.brainres.2013.03.024

Grafton, S.T., Arbib M.A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. *Experimental Brain Research, 112*, 103–11. http://dx.doi.org/10.1007/BF00227183

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley & Sons, Ltd, New York, NY, USA.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders, 39*, 350–365. http://dx.doi.org/10.1016/j.jcomdis.2006.06.013

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*, 280-301. http://dx.doi.org/10.1016/j.bandl.2005.06.001

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105*, 611–633. http://dx.doi.org/10.1037/0033-295X.105.4.611-633

Halle, M., & Stevens, K. N. (1959). Analysis by synthesis. In W. Wathen-Dunn & L. E. Woods (Eds.), *Proceedings of the seminar on speech compression and processing*. USAF Camb. Res. Ctr. 2: Paper D7.

Hermansky, H. (1998). Should recognizers have ears? *Speech Communication, 25*, 3-27.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience. 13*, 135-145. doi:10.1038/nrn3158

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., A. Senior, Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Magazine, 29*,

82-97.

Hochmann, J.R., & Papeo, L. (2014). The Invariance Problem in Infancy: A Pupillometry Study. *Psychological Science*, *25*, 2038–2046. http://dx.doi.org/10.1177/0956797614547918

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*. doi:10.3389/fnhum.2011.00082

Huang, X., & Deng, L. (2010). An Overview of Modern Speech Recognition. , in R. Herbrich & T. Graepel (Eds.) *Handbook of Natural Language Processing*, Second Edition, pp. 339-366. Chapman & Hall/CRC.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C., & Rizzolatti, G. (1999). Cortical Mechanisms of Human Imitation. *Science, 286*, 2526–2528. http://dx.doi.org/10.1126/science.286.5449.2526.

Imada T, Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P.K. (2006) Infant speech perception activates Broca's area: A developmental magnetoencephalography study. *NeuroReport, 17*, 957–962.

Ito, T., Tiede, M., & Ostry, D.J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences, 106*, 1245–1248. http://dx.doi.org/10.1073/pnas.0810063106.

Jacquemot, C., Dupoux, E., & Bachoud-Lévi, A.-C. (2007). Breaking the mirror: asymmetrical disconnection between the phonological input and output codes. *Cognitive Neuropsychology, 24*, 3–22. http://dx.doi.org/10.1080/02643290600683342

Jones, J.A., & Callan, D.E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport, 14,* 1129–1133. http://dx.doi.org/10.1097/00001756-200306110-00006

Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology, 23*, 648–654. http://dx.doi.org/10.1037/0012-1649.23.5.648

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73*, 322–335. http://dx.doi.org/10.1121/1.388813

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2009). Perception and hierarchical dynamics. *Frontiers in Neuroinformatics*, 3, 20. http://dx.doi.org/10.3389/neuro.11.020.2009

Kingston, J., & Diehl, R.L. (1994). Phonetic knowledge. *Language, 70*, 419–54. http://dx.doi.org/10.1353/lan.1994.0023.

Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–19. Association for Computational Linguistics.

Kleinschmidt, D. & Jaeger, T. F. (2015). Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological Review, 122,* 148–203. http://dx.doi.org/10.1037/a0038695

Kluender KR. (1994). Speech perception as a tractable problem in cognitive science. In M.A. Gernsbacher (Ed.) *Handbook of Psycholinguistics* (pp. 173–217). San Diego, CA: Academic.

Kröger, B.J., Kannampuzha, J., & Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics, 2*. http://dx.doi.org/10.1140/epjnbp15

Kröger, B.J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication, 51*, 793–809. http://dx.doi.org/10.1016/j.specom.2008.08.002

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America, 66*, 1668–1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior & Development, 6*, 263–285.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5, 831–843. http://dx.doi.org/10.1038/nrn1533

Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, T., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*, 979–1000. http://dx.doi.org/10.1098/rstb.2007.2154

Kuhl, P. K., & Meltzoff (1996). A. Infant vocalizations in response to speech: vocal imitation and developmental change. *Journal of the Acoustical Society of America*. *100*, 2425–2438.

Kuhl, P.K., Ramírez, R.R., Bosseler, A., Lotus Lin, J.F., & Imada, T. (2014). Infants' brain responses to speech suggest Analysis by Synthesis. *Proceedings of the National Academy of Sciences*, *111*, 12572–12573. http://dx.doi.org/10.1073/pnas.1410963111

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*, 606–608. http://dx.doi.org/10.1126/science.1736364

Laurent, R., Schwartz, J.-L., Bessière, P., & Diard, J. (2013). A computational model of perceptuo-motor processing in speech perception: learning to imitate and categorize synthetic CV syllables. In *Proceedings of InterSpeech* (pp. 2797–2801). Lyon, France.

Lebeltel, O., Bessière, P., Diard, J., & Mazer, E. (2004). Bayesian robot programming. *Autonomous Robots*, *16*, 49–79. http://dx.doi.org/10.1023/B:AURO.0000008671.38949.43

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of speech code. *Psychological Review, 74*, 431–461. http://dx.doi.org/10.1037/h0020279

Liberman, A.M., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36. http://dx.doi.org/10.1016/0010-0277(85)90021-6

Liberman, A.M., & Mattingly, I.G. (1989). A specialization for speech perception. *Science, 243*, 489–494. http://dx.doi.org/10.1126/science.2643163

1756 Liberman, A.M., & Whalen, D.H. (2000). On the relation of speech to language. *Trends in Cognitive*
1757 *Sciences, 4*, 187–196. http://dx.doi.org/10.1016/S1364-6613(00)01471-6
1758 Lotto, A.J. (2000). Language acquisition as complex category formation. *Phonetica, 57*, 189–96.
1759 http://dx.doi.org/10.1159/000028472
1760 Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood
1761 activation, and PARSYN. *Perception and Psychophysics, 62*, 615–625.
1762 MacNeilage, P.F. (1998). The frame/content theory of evolution of speech production. *Behavioral*
1763 *and Brain Sciences, 21*, 499–546. http://dx.doi.org/10.1017/S0140525X98001265
1764 Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis
1765 of vocal tract shapes using an articulatory model. In W. Hardcastle & A. Marchal (Eds.),
1766 *Speech production and speech modeling* (pp. 131–149). Kluwer Academic.
1767 Maeda, S. & Honda, K. (1994). From EMG to formant patterns: the implication of vowel spaces.
1768 *Phonetica*, *51*, 17–29. http://dx.doi.org/10.1159/000261955
1769 Marr, D. (1982). *Vision. A Computational Investigation into the Human Representation and*
1770 *Processing of Visual Information*. W.H. Freeman and Company, New York, USA.
1771 Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*.
1772 Laurence Erlbaum Associates, London.
1773 Massaro, D. W. (1998). *Perceiving Talking Faces*. MIT, Cambridge, Ma).
1774 Massaro, D.W., & Oden, G.C. (1980). Evaluation and integration of acoustic features in speech
1775 perception. *Journal of the Acoustical Society of America, 67*, 996–1013.
1776 http://dx.doi.org/10.1121/1.383941
1777 McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive*
1778 *Psychology, 18*, 1–86. http://dx.doi.org/10.1016/0010-0285(86)90015-0
1779 McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
1780 http://dx.doi.org/doi:10.1038/264746a0

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental science*, *12*, 369–378. http://dx.doi.org/10.1111/j.1467-7687.2009.00822.x

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*, 143-178.

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology, 17*, 1692–1696. http://dx.doi.org/10.1016/j.cub.2007.08.064

Meltzoff, A.N., & Moore, M.K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 4312*, 75–78. http://dx.doi.org/10.1126/science.198.4312.75

Ménard, L., Schwartz, J.-L., Boë, L.-J., Kandel, S., & Vallée, N. (2002). Auditory normalization: of French vowels synthesized by an articulatory model simulating growth from birth to adulthood. *Journal of the Acoustical Society of America*, *111*, 1892–1905. http://dx.doi.org/10.1121/1.1459467

Ménard, L., Schwartz, J.-L., & Aubin, J. (2008). Invariance and variability in the production of the height feature in French vowels. *Speech Communication, 50*, 14–28. http://dx.doi.org/10.1016/j.specom.2007.06.004

Ménard, L., Schwartz, J.-L., & Boë, L.-J. (2004). The role of vocal tract morphology in speech development: Perceptual targets and sensori-motor maps for French synthesized vowels from birth to adulthood. *Journal of Speech Language and Hearing Research, 47*, 1059–1080. http://dx.doi.org/10.1044/1092-4388(2004/079)

Meunier, C. (2007). Phonétique acoustique. In P. Auzou (Ed.), *Les dysarthries* (pp. 164–173). Solal.

Moore, R. (2007). Spoken language processing: Piecing together the puzzle. *Speech Communication, 49*, 418–435. http://dx.doi.org/10.1016/j.specom.2007.01.011

Möttönen, R., Dutton, R., & Watkins, K.E. (2013). Auditory-motor processing of speech sounds.

Cerebral Cortex, 23, 1190–1197. http://dx.doi.org/10.1093/cercor/bhs110

Möttönen, R., & Watkins, K.E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *The Journal of Neuroscience, 29*, 9819–9825. http://dx.doi.org/10.1523/JNEUROSCI.6018-08.2009

Moulin-Frier, C., Diard, J., Schwartz, J.-L., & Bessière, P. (2015). COSMO ("Communicating about Objects using Sensory-Motor Operations"): a Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics*, 53, 5-41.

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L, & Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor, and perceptuo-motor theories of speech perception: An exploratory Bayesian modelling study. *Language and Cognitive Processes, 27*, 1240–1263. http://dx.doi.org/10.1080/01690965.2011.645313

Moulin-Frier, C., Schwartz, J.-L., Diard, J., & Bessière, P. (2011). Emergence of phonology through deictic games within a society of sensori-motor agents in interaction. In A. Vilain, J.-L. Schwartz, C. Abry & J. Vauclair (Eds.) *Primate Communication and Human Language* (pp. 193–220). John Benjamins Publishing Company.

Nearey, T.M. (1990). The segment as a unit of speech perception. *Journal of Phonetics, 18*, 347–73

Norris D., & McQueen J.M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychol Rev., 115*, 357-95. doi: 10.1037/0033-295X.115.2.357.

Öhman, S. (1966). Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151–168. http://dx.doi.org/10.1121/1.1909864

Ojanen, V., Möttönen, R., Pekkola, Jääskeläinen, I., Joensuu, R., Autti, T., & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage, 25*, 333–338. http://dx.doi.org/10.1016/j.neuroimage.2004.12.001.

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N. E. et al. (2012).

Reconstructing speech from human auditory cortex. *PLoS Biol*. 10:e1001251. doi:10.1371/journal.pbio.1001251

Patri, J.F. Diard, J., & Perrier, P. (2015). Optimal speech motor control and token-to-token variability: a Bayesian modeling approach. *Biological Cybernetics, 109*, 611-626. http://link.springer.com/article/10.1007/s00422-015-0664-4.

Perkell, J. & Klatt, D. H. (1986). *Invariance and variability in speech processes*. Erlbaum.

Perrier, P. (2005). Control and representations in speech production. *ZAS Papers in Linguistics, 40*, 109–132.

Perrier, P., Lœvenbruck, H., & Payan, Y. (1996). Control of tongue movements in speech: The equilibrium point hypothesis perspective. *Journal of Phonetics, 24*, 53–75.

Pitermann, M. (2000). Effect of speaking rate and contrastive stress on formant dynamics and vowel perception. *Journal of the Acoustical Society of America, 107*, 3425–3437. http://dx.doi.org/10.1121/1.429413

Polka, L., Masapollo, M., & Ménard, L. (2014). Who's talking now? Infants' perception of vowels with infant vocal properties. *Psychological Science, 25*, 1448–1456.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences, 103*, 7865–7870. http://dx.doi.org/10.1073/pnas.0509989103

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996a). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131–141. http://dx.doi.org/10.1016/0926-6410(95)00038-0

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996b). Localization of grasp representations in humans by PET: 1. observation vs. execution. *Experimental Brain Research, 111*, 246–252. http://dx.doi.org/10.1007/BF00227301

Robert-Ribes, J., Schwartz, J.-L., Lallouache, T. & Escudier, P. (1998). Complementarity and synergy

in bimodal speech : auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America*, 103, 3677-3689. http://dx.doi.org/10.1121/1.423069

Rogers, J. C., Möttönen, R., Boyles, R., & Watkins, K. E. (2014). Discrimination of speech and non-speech sounds following theta-burst stimulation of the motor cortex. *Frontiers in Psychology, 5*, 754. http://dx.doi.org/10.3389/fpsyg.2014.00754

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, 12, 573–604. http://dx.doi.org/10.3758/BF03196750

Sato, M., Grabski, K., Glenberg, A., Brisebois, A., Basirat, A., Ménard, L. & Cattaneo, L. (2011). Articulatory bias in speech categorization: evidence from use-induced motor plasticity. *Cortex, 47*, 1001–1003. http://dx.doi.org/10.1016/j.cortex.2011.03.009

Sato, M., Tremblay, P. & Gracco, V. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain and Language, 111*, 1–7. http://dx.doi.org/10.1016/j.bandl.2009.03.002

Savariaux, C., Perrier P., & Orliaguet, J.-P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: a study of the control space in speech production. *Journal of the Acoustical Society of America, 98*, 2428–2442. http://dx.doi.org/10.1121/1.413277

Savariaux, C., Perrier, P., Orliaguet, J.P. & Schwartz, J.-L. (1999). Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis. *Journal of the Acoustical Society of America, 106*, 381–393. http://dx.doi.org/10.1121/1.427063

Scharenborg, O., Norris, D., Ten Bosch, L., & McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science, 29*, 867-918. doi:10.1207/s15516709cog0000_37.

Schwartz, J.L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *Journal of the Acoustical Society of America*, *127*, 1584-

1594. http://dx.doi.org/10.1121/1.3293001

Schwartz, J.-L., Abry, C., Boë, L.-J., & Cathiard, M. (2002). Phonology in a theory of perception-for-action-control. In J. Durand, B. Laks (Eds.) *Phonology: from Phonetics to Cognition* (pp. 255–280). Oxford: Oxford University Press.

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012a). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics, 25, 336–354*. doi:10.1016/j.jneuroling.2009.12.004

Schwartz, J.-L., Boë, L.-J., & Abry, C. (2007). Linking the Dispersion-Focalization Theory (DFT) and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a Perception-for-Action-Control Theory (PACT). In M.J. Solé, P. Beddor & M. Ohala (Eds.) *Experimental Approaches to Phonology* (pp. 104–124). Oxford University Press.

Schwartz J.-L., Boë L.-J., Badin P., & Sawallis R. T. (2012b). Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial-coronal-velar stop series, *Journal of Phonetics, 40*, 20–36.

Schwartz, J.L., Boë, L.J., Vallée, N., & Abry , C. (1997). Major trends in vowel system inventories. *Journal of Phonetics, 25*, 233-254.

Shiller, D.M., Sato, M., Gracco, V.L., & Baum, S.R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *Journal of the Acoustical Society of America, 125,* 1103–1113. http://dx.doi.org/10.1121/1.3058638

Skipper, J.I., Devlin, J.T., & Lametti, D.R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language, 164*, 77-105. http://dx.doi.org/10.1016/j.bandl.2016.10.004

Skipper, J.I., van Wassenhove, V., Nusbaum, H.C. & Small, S.L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17*, 2387–2399. http://dx.doi.org/10.1093/cercor/bhl147.

Sonderegger, M., & Yu, A. (2010). A rational account of perceptual compensation for coarticulation. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pp. 375–380.

Stevens, K.N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In E. David & P. Denes (Eds.), *Human communication: a unified view* (pp. 51–66). McGraw-Hill.

Stevens, K.N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America, 68*, 836–842. http://dx.doi.org/10.1121/1.384823

Stevens, K.N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3-45.

Stevens, K.N., Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64*, 1358–1368. http://dx.doi.org/10.1121/1.382102

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form* (pp. 88–102). Cambridge, MA: MIT Press.

Strange, W. (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America, 85*, 2081–2087. http://dx.doi.org/10.1121/1.397860

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212–215. http://dx.doi.org/doi: 10.1121/1.1907309

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B Dodd and R Campbell (Eds.) *Hearing by eye: the psychology of lip-reading* (pp 3–51). Lawrence Erlbaum Associates, London.

Sun, J., & Deng, L. (2002). An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition. *Journal of the Acoustical Society of America, 111*, 1086–1101.

Sussman, H., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: the orderly output constraint. *Behavioral and Brain Sciences*, 21(02), 241–259.

1931 Sussman, H. M., Hoemeke, K. & Ahmed, F. (1993) A cross-linguistic investigation of locus equations
1932 as a relationally invariant descriptor for place of articulation. *Journal of the Acoustical Society*
1933 *of America, 94*, 1256–68. http://dx.doi.org/10.1121/1.408178
1934 Sussman, H. M., McCaffrey, H. A. & Matthews, S. A. (1991) An investigation of locus equations as a
1935 source of relational invariance for stop place categorization. *Journal of the Acoustical Society*
1936 *of America, 90*, 1309–25. http://dx.doi.org/10.1121/1.401923
1937 Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological*
1938 *Review*, *61*, 401–409. http://dx.doi.org/10.1037/h0058700
1939 Toscano, J. C., & McMurray, B. (2008). Using the distributional statistics of speech sounds for
1940 weighting and integrating acoustic cues. In *Proceedings of the 30th Annual Conference of the*
1941 *Cognitive Science Society*, pp. 433–439.
1942 Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in
1943 speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34, 434–
1944 464. http://dx.doi.org/10.1111/j.1551-6709.2009.01077.x
1945 Treille, A., Cordeboeuf, C., Vilain, C., & Sato, M. (2014). Haptic and visual information speed up the
1946 neural processing of auditory speech in live dyadic interactions. *Neuropsychologia*, 57, 71–77.
1947 http://dx.doi.org/doi:10.1016/j.neuropsychologia.2014.02.004
1948 Vallabha, G. K., McClelland, J. L., Pons F., Werker, J. F., & Amano, S. (2007). Unsupervised
1949 learning of vowel categories from infant-directed speech. *Proceedings of the National*
1950 *Academy of Sciences*, 104, 13273–13278. http://dx.doi.org/10.1073/pnas.0705369104
1951 Watkins, K.E., Strafella, A.P., & Paus, T. (2003). Seeing and hearing speech excites the motor system
1952 involved in speech production. *Neuropsychologia*, 41, 989–994.
1953 http://dx.doi.org/10.1016/S0028-3932(02)00316-0
1954 Werker, J.F., & Hensch, T.K. (2015). Critical Periods in Speech Perception: New Directions. *Annu.*
1955 *Rev. Psychol.. 66,* 173–96.

Werker, J.F., & Tees, R.C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63. http://dx.doi.org/10.1016/S0163-6383(84)80022-3

Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in productibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage, 33*, 316–325. http://dx.doi.org/10.1016/j.neuroimage.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*, 701–702. http://dx.doi.org/10.1038/nn1263

Zekveld, A.A., Heslenfeld, D.J., Festen, J.M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *NeuroImage, 32*, 1826–1836. http://dx.doi.org/10.1016/j.neuroimage.2006.04.199

Footnotes

(1) It is well known that other sensory systems may intervene in speech perception: such as vision, through lip-reading (Sumby & Pollack, 1954; McGurk & MacDonald, 1976; Summerfield, 1987) but also possibly somato-sensory processing (Fowler & Dekle, 1991; Ito, Tiede & Ostry, 2009; Treille et al., 2014). However, in this paper we will focus on the auditory vs. motor systems, acknowledging that these other sensory systems could be incorporated as additional sensory inputs inside a perceptuo-motor framework.

(2) Here, we use the term "communication object" in a broad sense, conflating different levels of analysis (phonetics, phonology, syntax, semantics). In this paper, objects will only refer to phonological entities.

(3) Technically, probability values below the ε threshold are set to ε during perception inference, but are set to 0 during learning. This makes learning approximate but fast, as portions of spaces with very low probabilities are dismissed altogether.

(4) Notice that the visual system could intervene in this process, especially considering the natural complementarity of auditory and visual representations in the depiction of vowels and plosives (Summerfield, 1987; Robert-Ribes et al., 1998).

(5) VLAM is actually an articulatory rather than a motor model of speech production. VLAM inputs are parameters controlling the shape of the tongue and lips and the position of the jaw, which are themselves the results of motor commands at a higher level (see e.g. Perrier et al., 1996; Perrier, 2005). We consider as a simplification that VLAM articulatory parameters are part of the control system and hence could provide "motor commands" at a certain level of representation in the motor pathway.

(6) While we display mean entropies in this Figure, averaging over the 9 syllables, there are actually differences between entropy dynamics among the different syllables, particularly in the motor space. This is clearly seen in Figure 14, where it appears that

convergence is more rapid for /u/ than for the other two vowels. The likely reason is that the available articulatory space for achieving the adequate formants is more restricted for /u/ in the available 3-D articulatory space in VLAM. Notice that slower articulatory convergence (displayed in Figure 14) can occur in spite of rapid acoustic convergence (as displayed by the rapid formant convergence for /i/ in Figure 14). It is beyond the scope of the present paper to discuss the importance and significance of these differences in convergence among vowels, plosives or syllables.

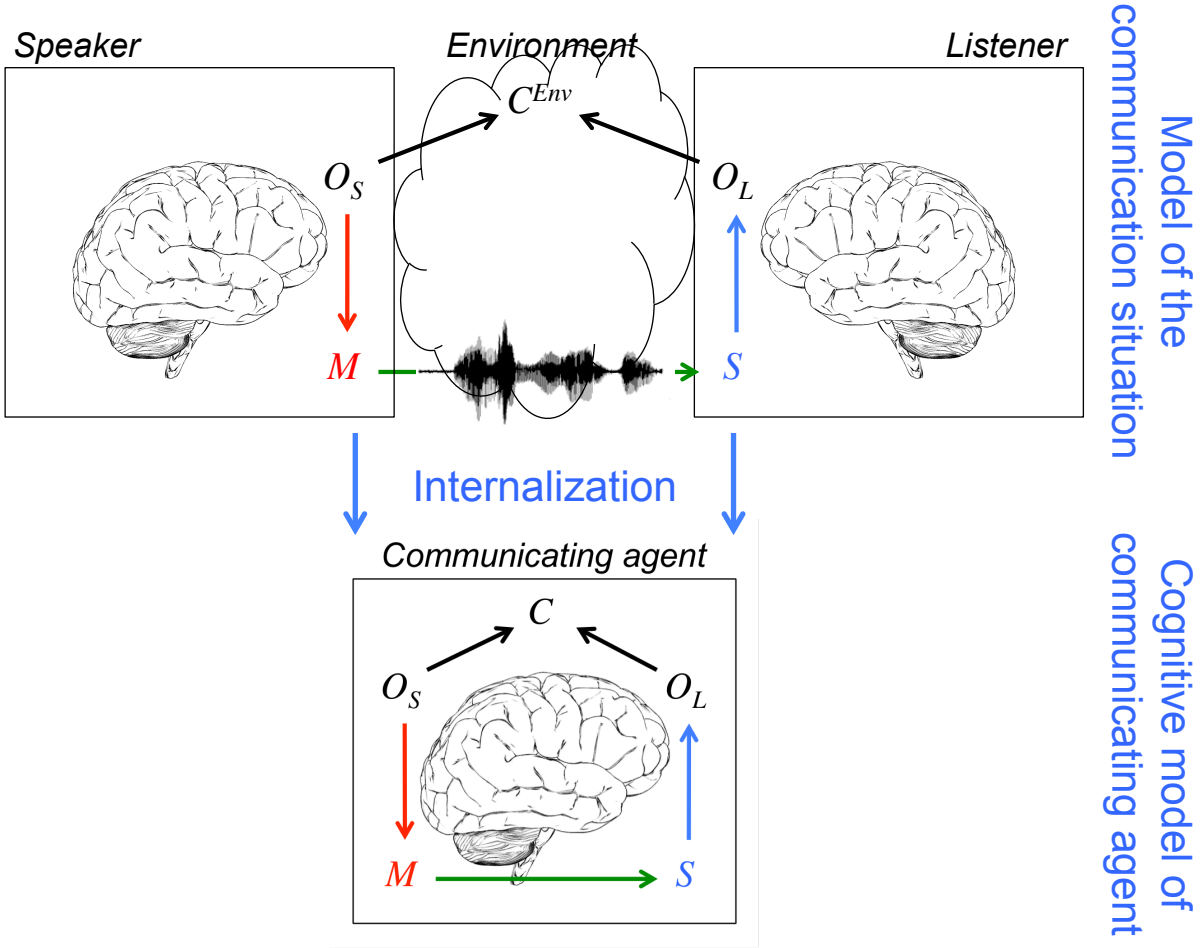


Figure 1. The communication situation, which involves a speaker agent and listener agents interacting within an environment, is internalized in communicating agents. Top, model of the communication situation: the speaker wants to mention a linguistic object (in a broad sense, see footnote 2) O_S. She/he produces a motor gesture M leading to the production of a sound S propagating in the environment towards the listeners who recover linguistic objects O_L. The success of communication is estimated by the Boolean variable C^{Env}. Bottom, all variables are internalized to provide a cognitive model of the communicating agent.

	Production infer $P(M \mid O)$	Perception infer $P(O \mid S)$
Motor theory focus on O_S	$\underbrace{P(M \mid O_S)}_{\text{motor repertoire}}$	$\propto \sum_M \Big(\underbrace{P(M \mid O_S)}_{\text{motor decoder}} \quad \underbrace{P(S \mid M)}_{\text{inverse model}} \Big)$
Auditory theory focus on O_L	$\propto P(M) \sum_S \Big(\underbrace{P(S \mid M)}_{\text{direct model}} \quad \underbrace{P(O_L \mid S)}_{\text{sensory targets}} \Big)$	$\underbrace{P(O_L \mid S)}_{\text{sensory classifier}}$
Perceptuo-motor theory $C{=}True$, i.e. $O_S = O_L$	$\underbrace{\propto P(M \mid [O_S = O_L])}_{\text{motor production}} \underbrace{\sum_S \big(P(S \mid M) P(O_L \mid S)\big)}_{\text{sensory production}}$	$\underbrace{\propto P([O_L = O_S] \mid S)}_{\text{sensory perception}} \underbrace{\sum_M \big(P(M \mid O_S) P(S \mid M)\big)}_{\text{motor perception}}$

Figure 2. Probabilistic inferences for production and perception tasks instantiated within the framework of the motor, auditory and perceptuo-motor theories. The ∝ symbol denotes proportionality, i.e. to correctly obtain probability distributions, the expression shown has to be normalized. The denominations of the components of each equation refer to their possible interpretation in terms of cognitive processes:

- sensory targets refer to the set of sensory distributions for each object (typically, "sensory" would be replaced by "auditory" in a basic auditory theory of speech perception, or possibly "audio-visual" in a modified version taking into account lip-reading: see note (1)),
- motor repertoire refers to the set of motor distributions for each object,
- sensory production refers to the distribution of sensory data (typically sounds) for each object,
- motor production refers to the distribution of motor commands (typically articulatory gestures) for each object,
- sensory classifier refers to the possibility of recovering the object from the stimulus input (typically the sound),
- motor decoder refers to the possibility (thanks to the Bayesian summation) of recovering the object from the motor commands (or, in some variants of motor

theories, the articulatory gesture),

- direct model refers to the possibility of predicting sensory information from the motor command (typically sound from the gesture),
- inverse model refers to the possibility of recovering the motor command from sensory information (typically the gesture from sound).

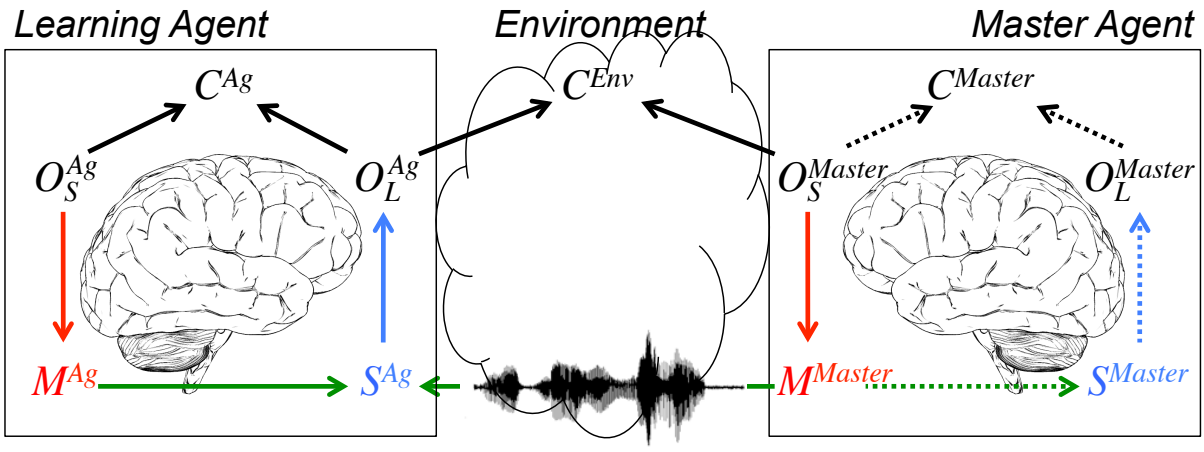
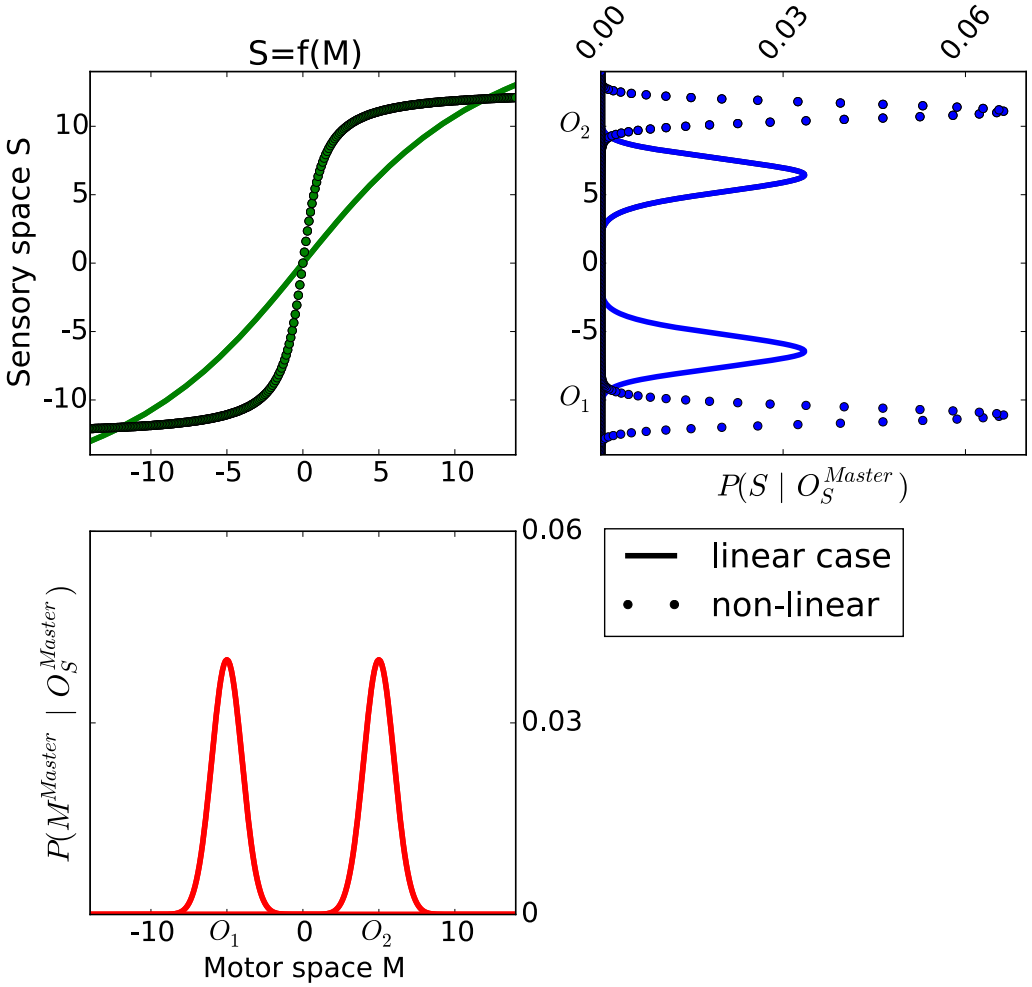


Figure 3. Schema of the supervised learning scenario, where the Master Agent provides the Learning Agent with <object, stimulus> pairs.



2044

2045 *Figure 4.* Summary of the stimulus production process of the Master Agent. Its motor

2046 repertoire is shown in the lower left panel. The model f of the motor-to-sensory

2047 transformation is shown in the upper left panel, for two values of the nonlinearity parameter

2048 ($a = 0.01$ for the quasi-linear case and $a = 0.1$ for the nonlinear case). The probability

2049 distributions of the resulting sensory inputs received by the Learning Agent are shown in the

2050 upper right panel.

2051

2052

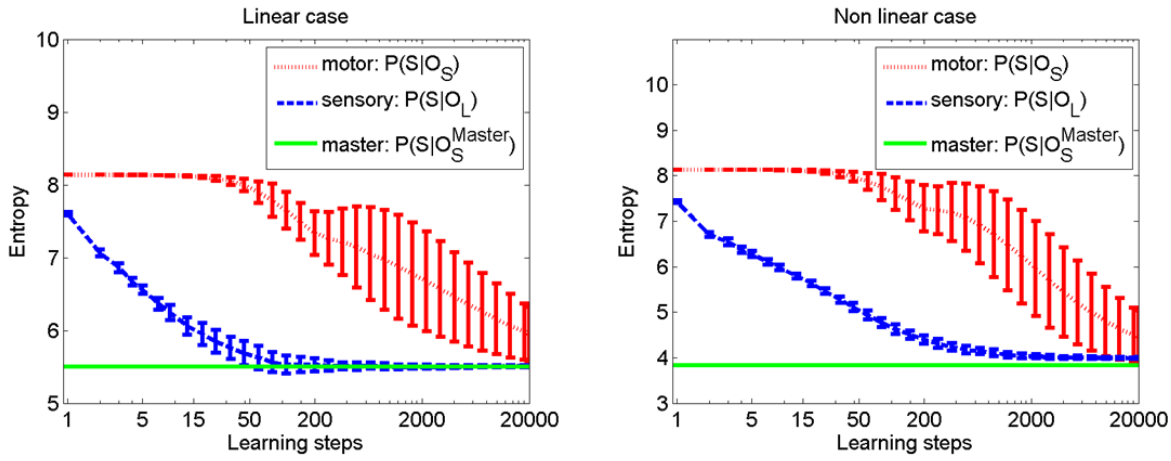


Figure 5. Evolution of entropies of the sensory and motor models of the Learning Agents and production system of the Master Agent, as a function of the number of iterations of the learning algorithm, averaged over the possible object values. Left column: linear case; right column: nonlinear case. In each case, 12 different simulations were run, corresponding to random initializations of the learning process. The standard deviations shown are computed over these 12 different simulations.

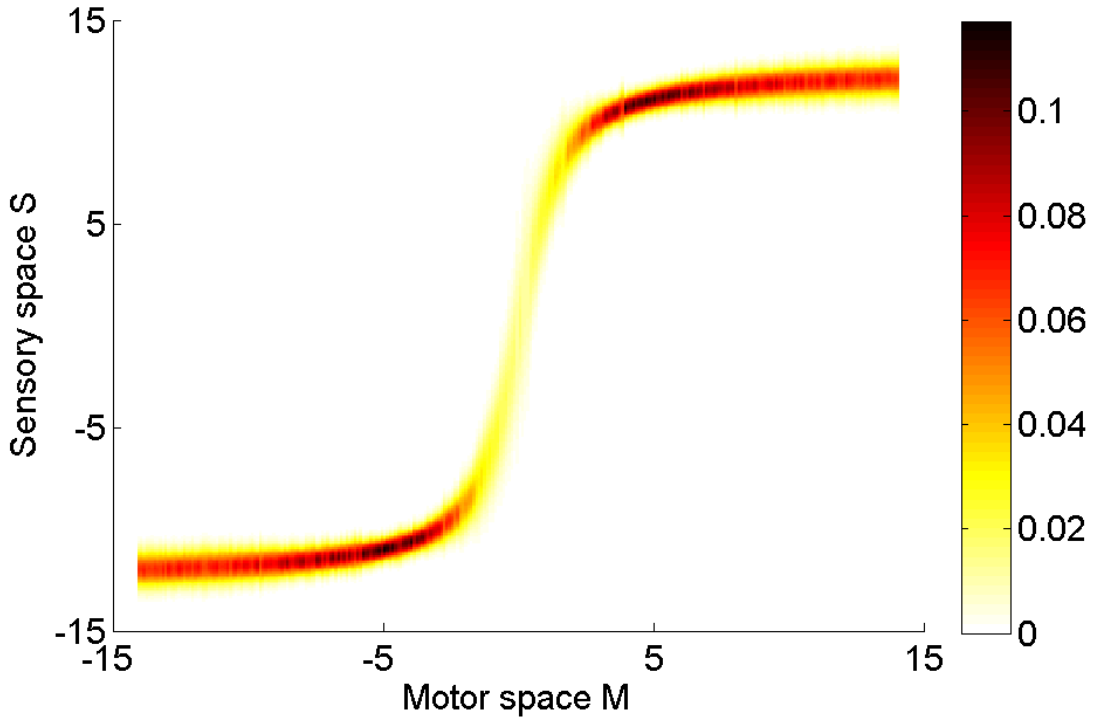


Figure 6. An instance of the learned internal model of the motor-to-sensory transformation, after 20,000 learning iterations, in a nonlinear setting ($a = 0.1$). For each motor gesture m of the x-axis, the probability distribution over resulting sensory stimulus $P(S \mid [M = m])$ is read vertically, with the color code indicating probability (white to yellow to red to black color-map (light gray to black), in order of increasing probability value). Black regions (resp. yellow/light gray) therefore correspond to low-variance (resp. high variance) Gaussian probability distributions, that is to say, well-explored (resp. poorly explored) portions of the motor-to-sensory transformation.

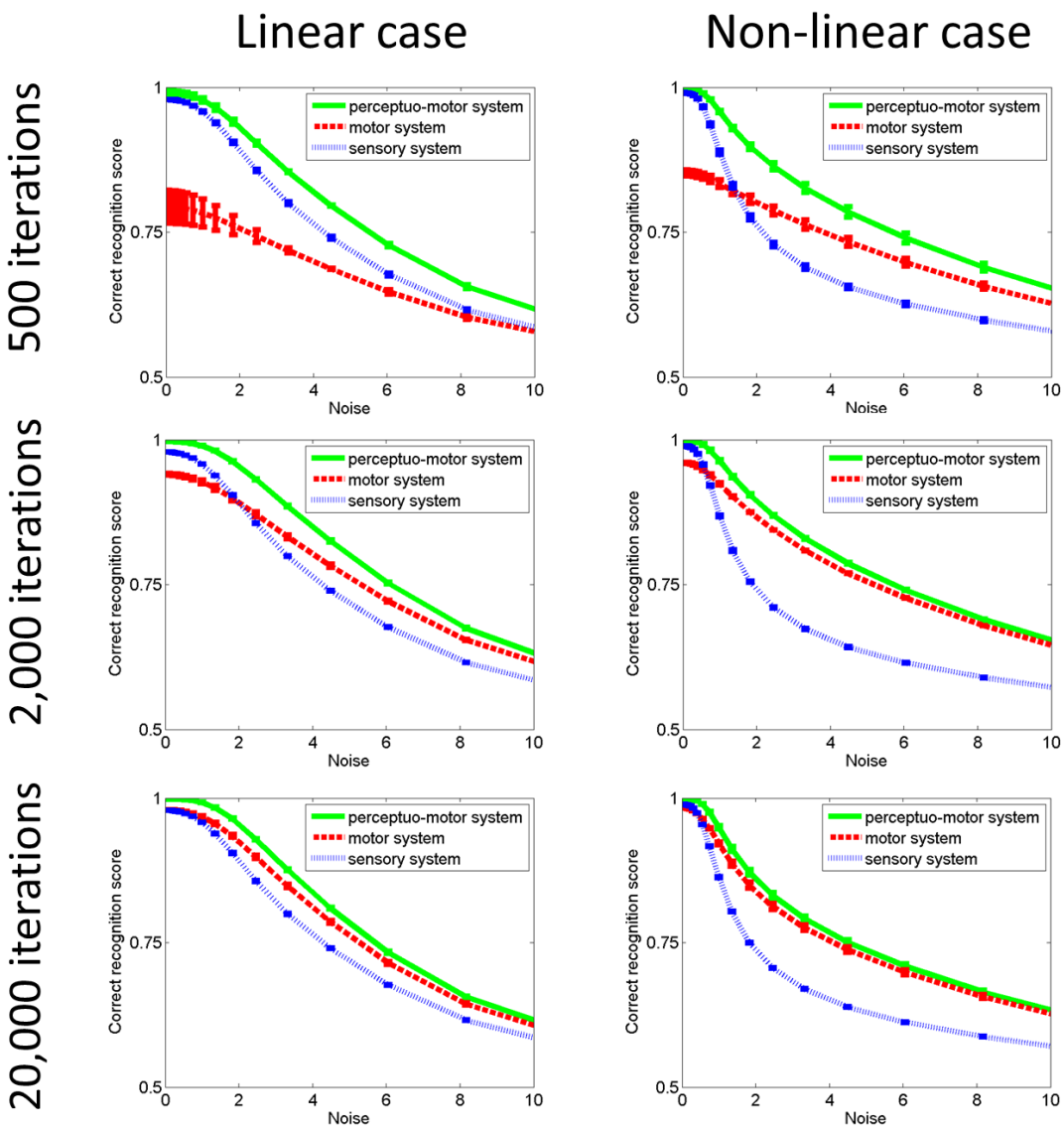


Figure 7. Evolution of correct recognition scores of motor, sensory and perceptuo-motor models of perception, as a function of environment noise. In each case, 12 different simulations were run, corresponding to random initializations of the learning process. The standard deviations shown are computed over these 12 different simulations.

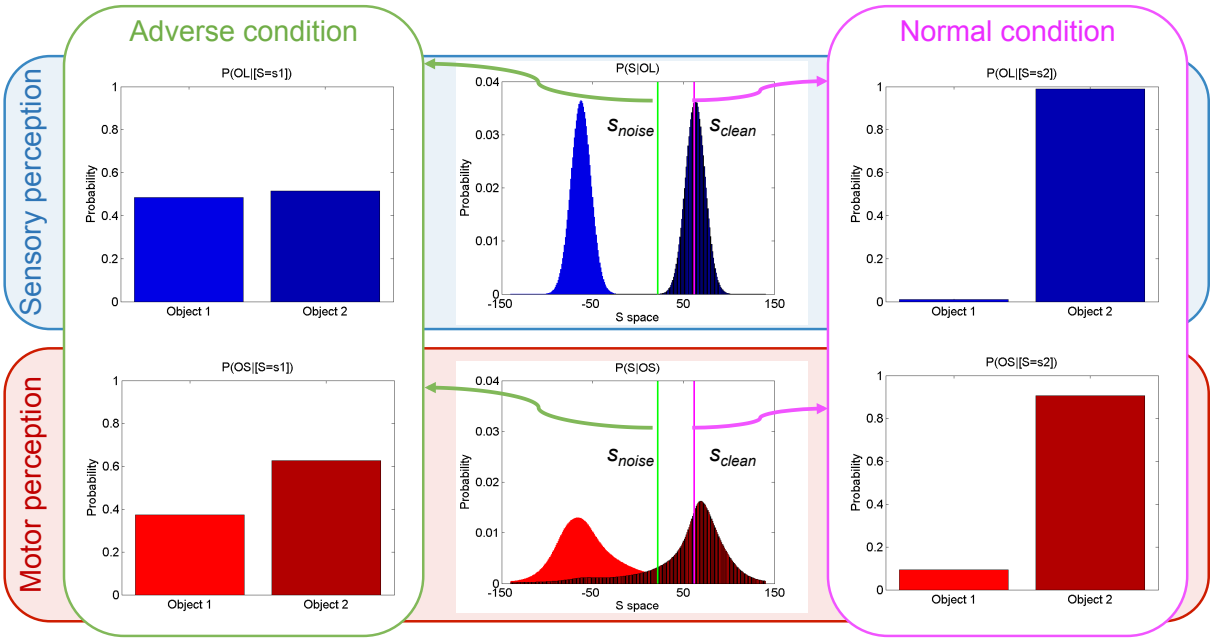
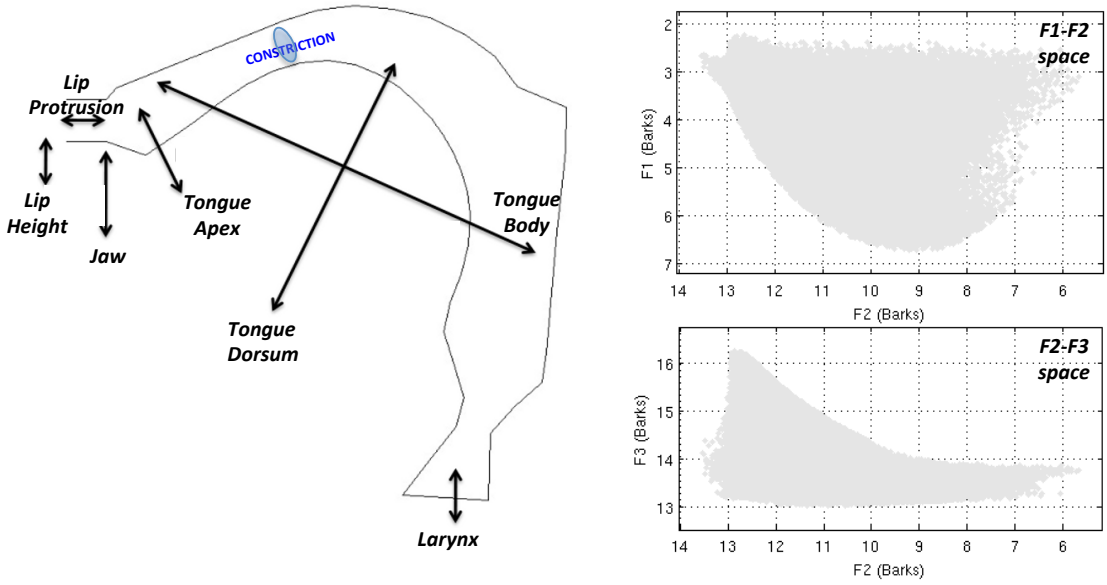


Figure 8. Illustration (linear case, after 1,200 learning iterations) of sensory (top row) and motor (bottom row) categorization processes on example stimuli s_{noise} in adverse conditions (left column) and s_{clean} in normal conditions (right column), as probabilistic inference from learned prototypes (center column).

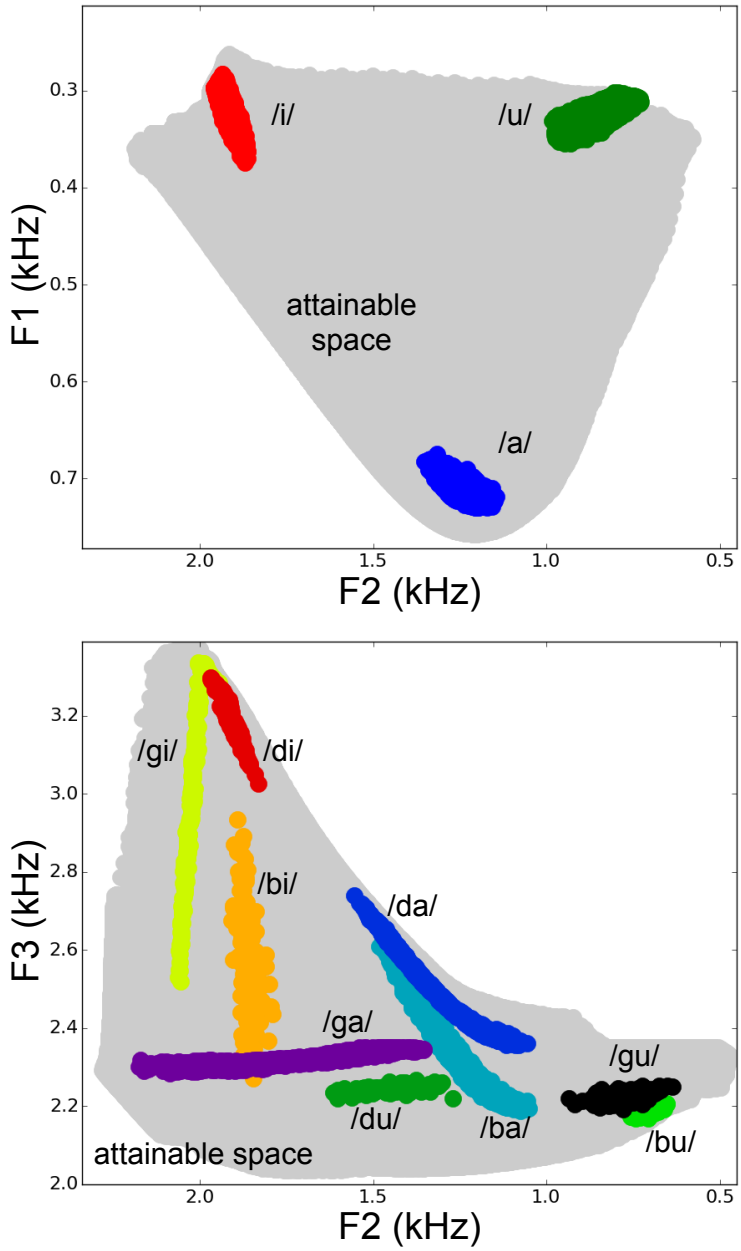
2082



2083

2084 *Figure 9.* The vocal tract VLAM model. Left: the seven articulatory parameters (Jaw, Lip
2085 Height and Protrusion, Tongue Body, Dorsum and Apex, and Larynx) enable the vocal tract
2086 shape to be driven. The Constriction is defined by the position where the vocal tract area is
2087 minimum. Vowels are constrained to have an area greater than 0.15 cm^2. Plosives are
2088 constrained to between 0.05 and 0.15 cm^2. Right: plots of the regions of the acoustic space
2089 (top: $(F1, F2)$ plane, bottom: $(F2, F3)$ plane) that result from articulatory configurations in
2090 VLAM.

2091



2092

2093 *Figure 10.* Synthetic syllables in acoustic space. Top: $(F2, F1)$ for vowels. Bottom: $(F2, F3)$

2094 for plosives in each syllabic context.

2095

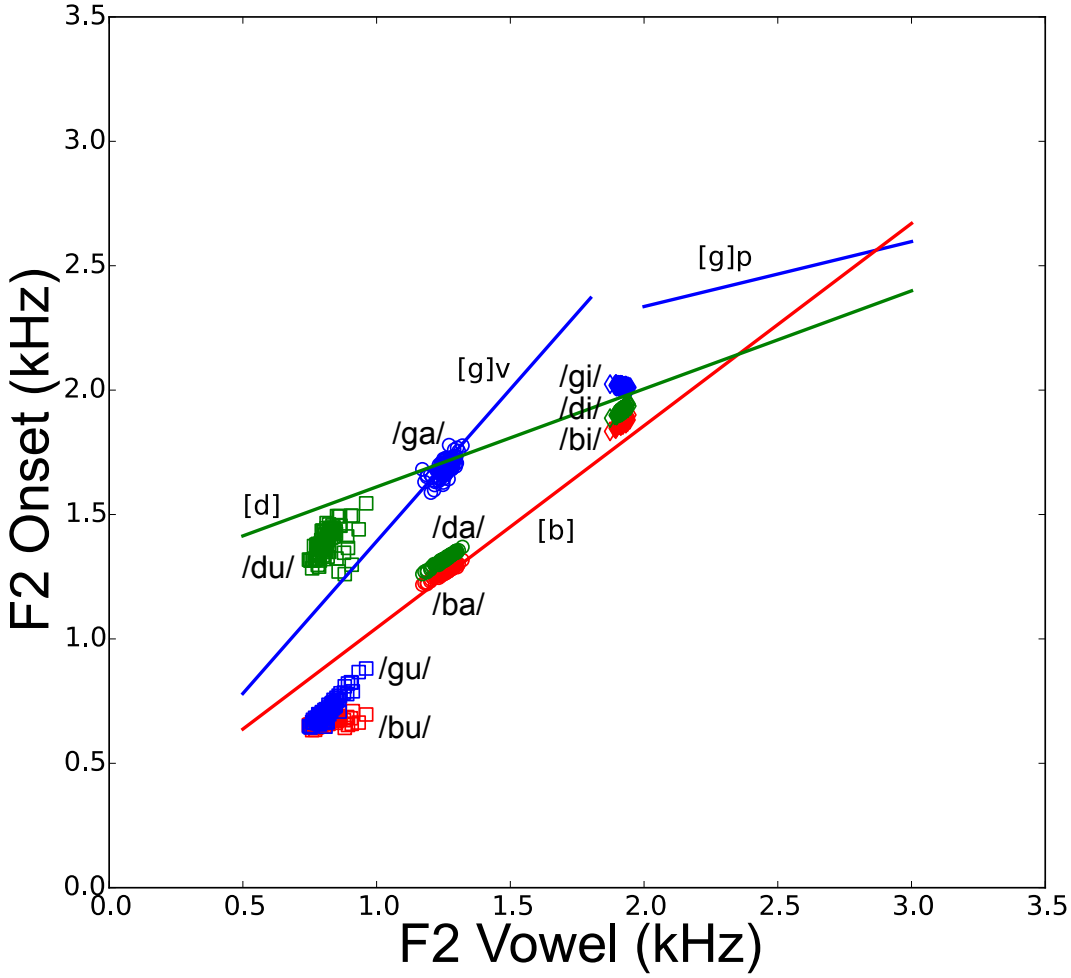
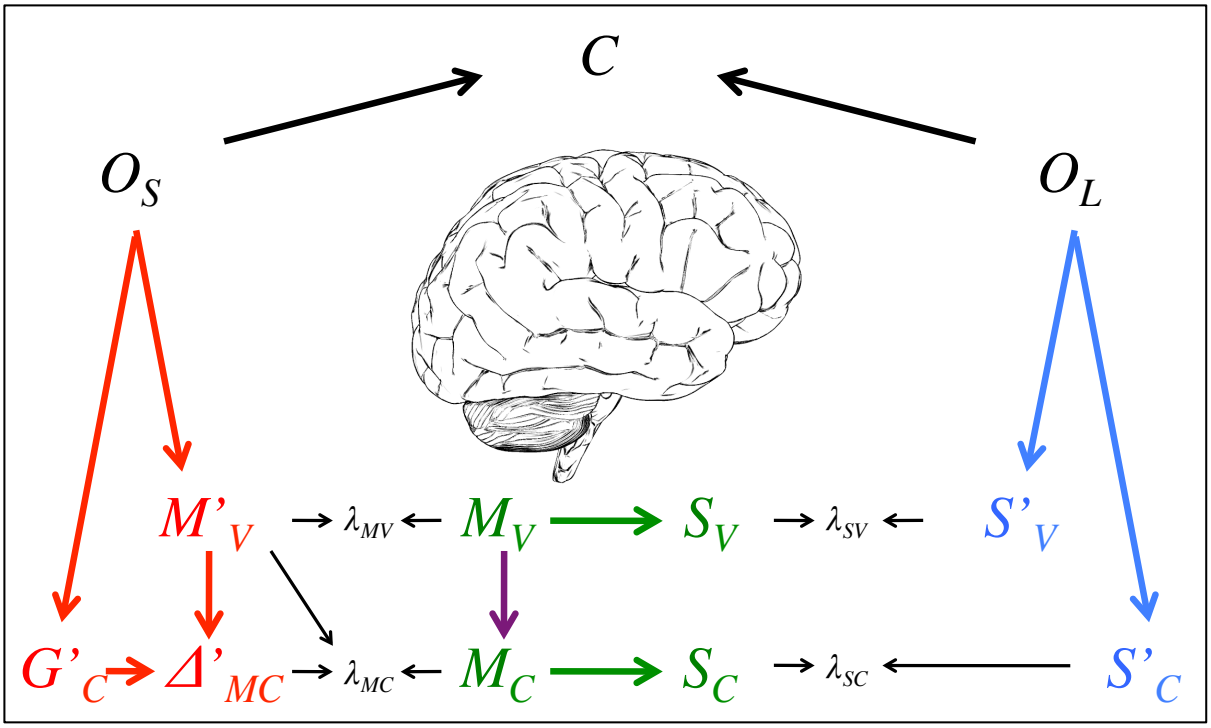


Figure 11. Locus displays for VLAM simulations (color marks annotated with /ba/, /bi/, /bu/, etc.) compared with locus equations provided by Sussman (1998) (portions of straight lines annotated with [d], [b], etc.). For VLAM simulations, each mark is displayed at the position corresponding to the *F*2 value for the vowel on the x-axis, and the *F*2 value for the consonant on the y-axis. Sussman's locus equations are derived by pooling the same frequency coordinates for natural utterances from 20 American English speakers (see Sussman, 1998, Fig. 5). Sussman provides one equation for /b/, one for /d/ and two separate equations for /g/: one when the context vowel is front and the plosive is therefore palatal, and another one when the context vowel is back and the plosive is therefore velar.



$$
\begin{aligned}
P(O_S\ G'_C\ M'_V\ \Delta'_{MC}\ \lambda_{MV}\ \lambda_{MC}\ M_V\ M_C\ S_V\ S_C\ \lambda_{SV}\ \lambda_{SC}\ S'_V\ S'_C\ O_L\ C) = \\
P(O_S)P(M'_V \mid O_S)P(G'_C \mid O_S)P(\Delta'_{MC} \mid M'_V\ G'_C) \\
P(\lambda_{MV} \mid M'_V\ M_V)P(\lambda_{MC} \mid M'_V\ \Delta'_{MC}\ M_C) \\
P(M_V)P(S_V \mid M_V)P(M_C \mid M_V)P(S_C \mid M_C) \\
P(\lambda_{SV} \mid S_V\ S'_V)P(\lambda_{SC} \mid S_C\ S'_C) \\
P(O_L)P(S'_V\ S'_C \mid O_L) \\
P(C \mid O_S\ O_L)
\end{aligned}
$$

Figure 12. The *COSMO-S* model for processing syllables. This is illustrated by a graphical representation (Top), and by the decomposition of its joint probability distribution as a product of probabilistic terms (Bottom). In red (left part), the motor system, in green (middle part), the sensory-motor system, in blue (right part), the auditory system.

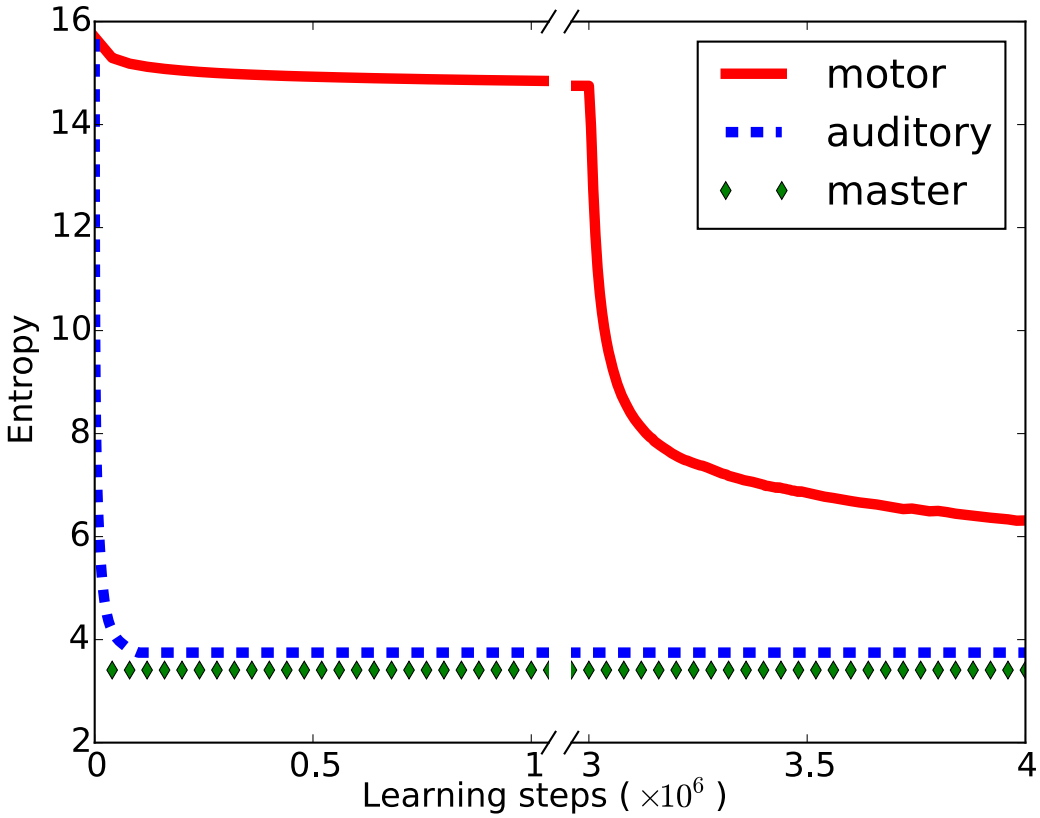


Figure 13. Evolution of entropies of the auditory and motor models of the Learning Agents and production system of the Master Agent, as a function of the number of iterations of the learning algorithm, averaged over the possible object values, in the syllable experiment. For the auditory model, learning corresponds to the sensory learning phase L1 with 4,000,000 iterations. For the motor model, learning starts with the sensory-motor learning phase L2 with 3,000,000 iterations followed by the motor learning phase L3 from 3,000,000 to 4,000,000 iterations.

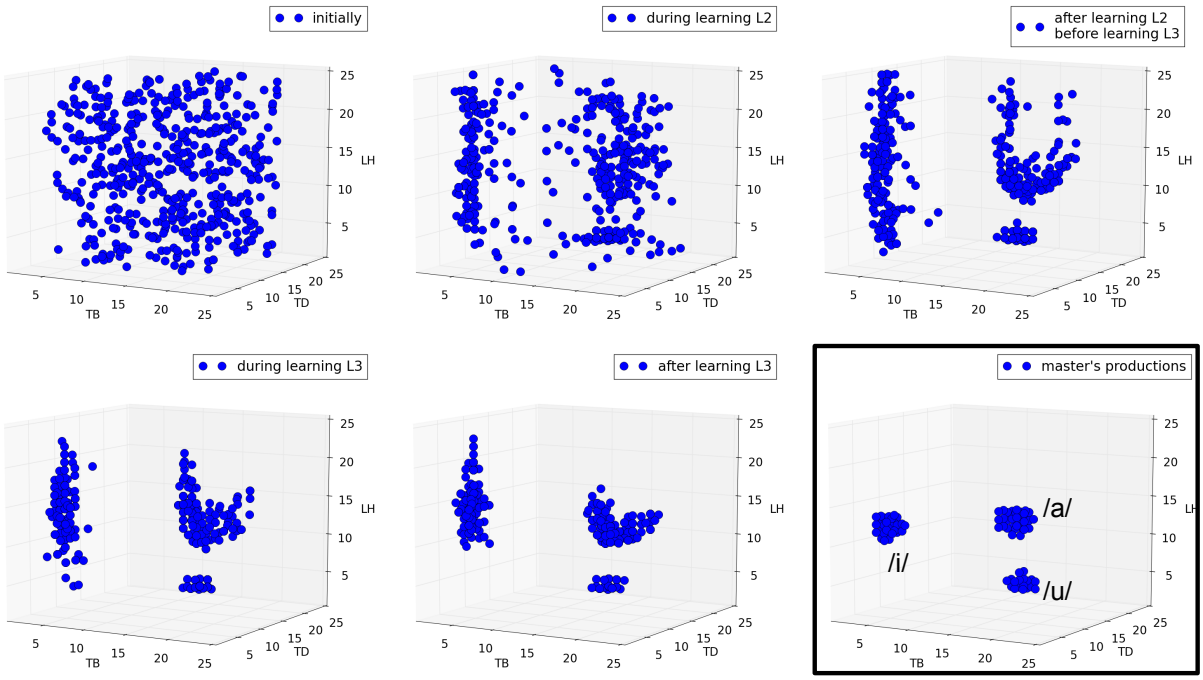


Figure 14. Illustrating exploration in the motor space in COSMO-S. Each graph displays samples from the probability distributions $P(M'^{Ag}_{V} \mid [S^{Ag}_{V} = s_v][O^{Ag}_{S} = o_S][\lambda^{Ag}_{MV} = 1])$ in the three-dimensional space *TB* (Tongue Body), *TD* (Tongue Dorsum) and *LH* (Lip Height), with s_v the vowel acoustic target and o_S the corresponding syllable label. Motor variables are specified by normalized values between 0 and 25. Each panel shows 500 samples taken at 500 successive time-steps (one sample per learning iteration), during five stages of the exploration process (see caption of each panel). The bottom right panel shows, for comparison, the motor distribution of the Master Agent.

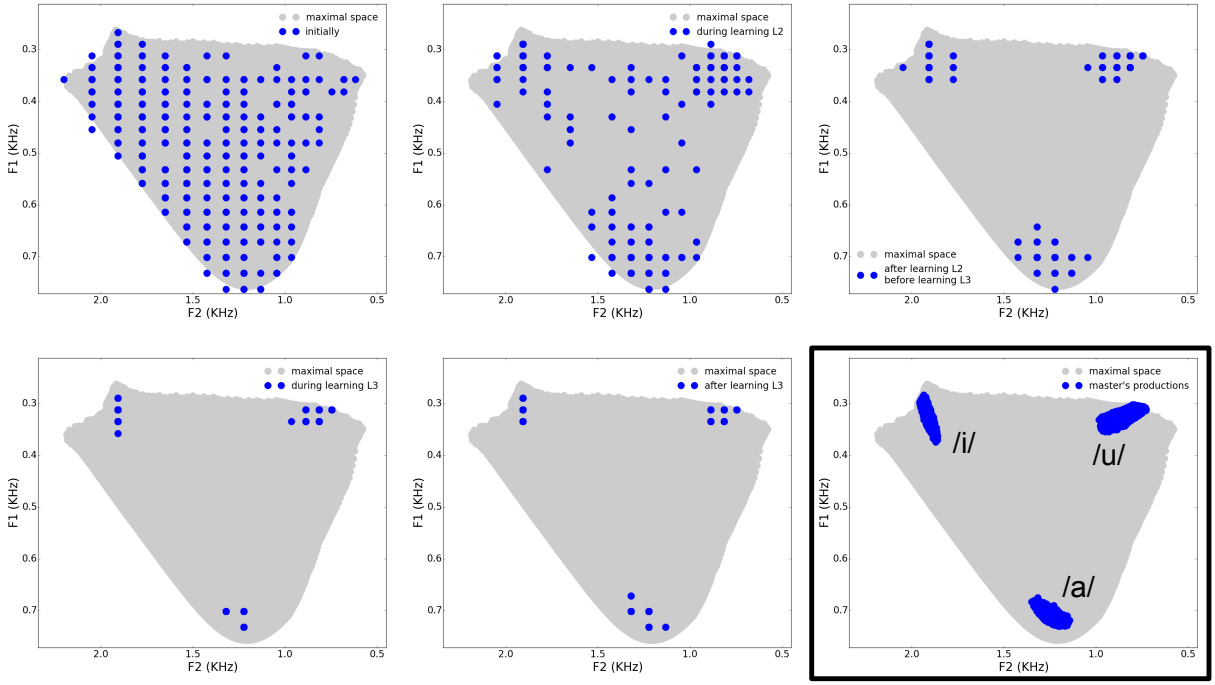


Figure 15. Illustrations of the exploration in the vowel space in COSMO-S. Each graph displays the images in the acoustic $(F2, F1)$ plane of the exact same motor samples as in Figure 14, via the articulatory-to-acoustic transformation. Each panel concerns the same five stages of the exploration process as in Figure 14. For comparison, the bottom right panel shows the stimulus distribution of the Master Agent.

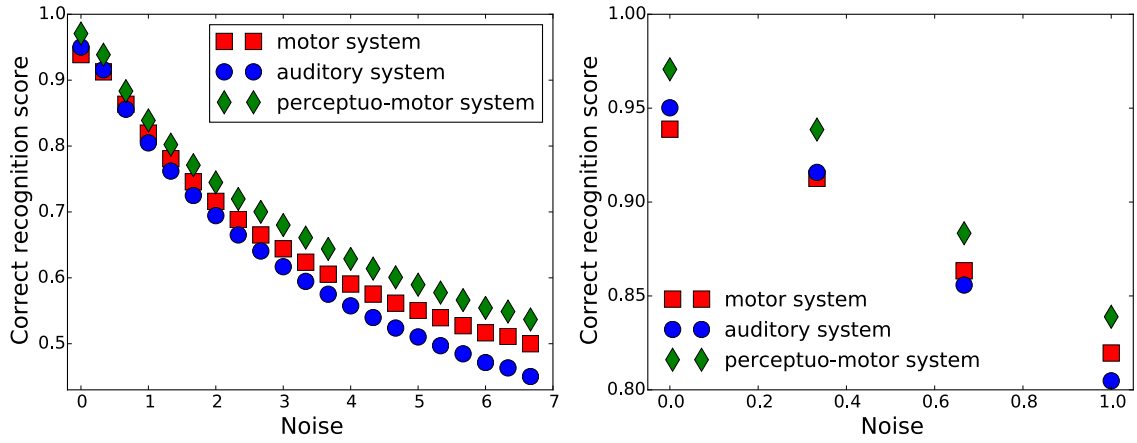


Figure 16. Results of the classification process for syllables presented at various levels of noise. The correct recognition rates for the auditory, motor and perceptuo-motor implementations of the perception task in the COSMO-S model are displayed. Right plot: zoom of the left plot at low levels of noise highlights the inversion of performance between the auditory system (better under normal conditions) and the motor system (better at noisy conditions).

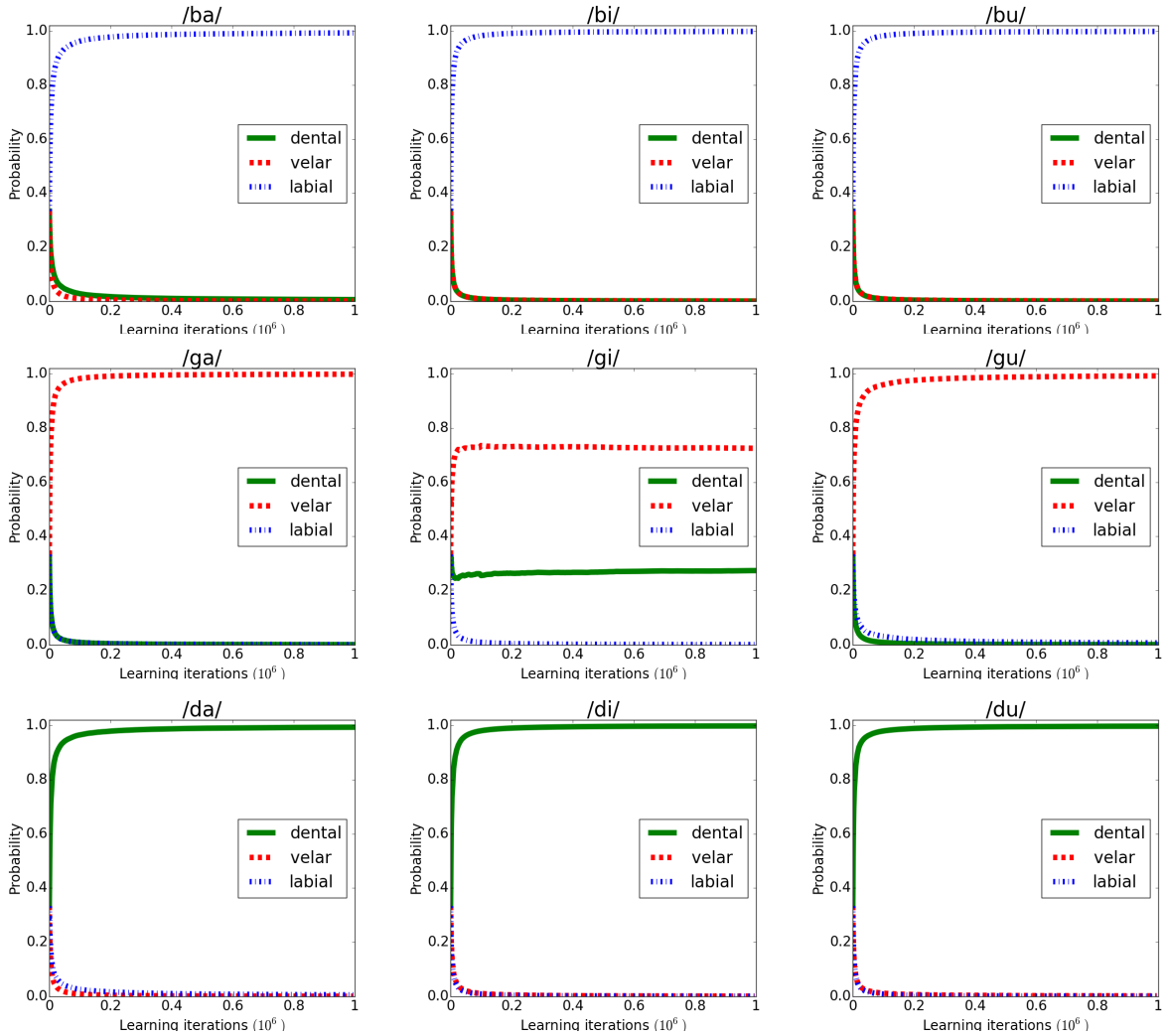


Figure 17. Learning the place of articulation for plosives. Evolution of the probabilities of the motor variable $P\left(G'^{Ag}_{C} \mid O^{Ag}_{S}\right)$ with the number of iterations in learning, for the 9 objects O^{Ag}_{S} (see text).

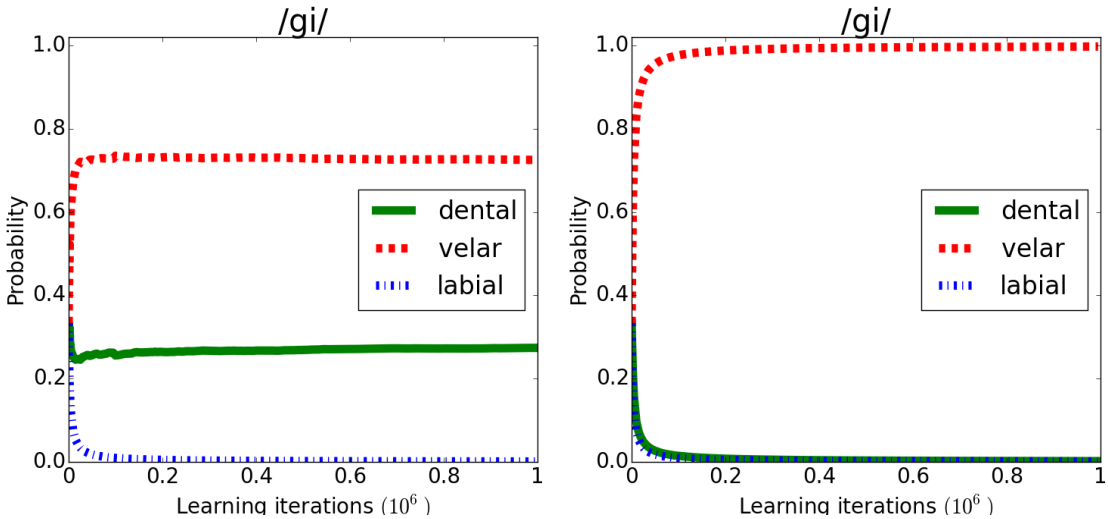


Figure 18. The role of hyperarticulation in the emergence of phonological categories. Evolution of the probabilities of the motor variable $P\left(G'^{Ag}_{C} \mid O^{Ag}_{S}\right)$ with the number of iterations in learning, for the object $O^{Ag}_{S} = /gi/$, comparing the cases where learning is without (left, identical to the middle panel of Figure 16) or with (right) hyper-articulation by the Master Agent (see text).